# Optimisation of the use of multiple sources of data in short-term photovoltaic generation forecasting models

Kevin Bellinguer

PSL
UNIVERSITÉ PARIS

**THÈSE DE DOCTORAT**
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

# Optimisation de l'Intégration de Données Multi-sources dans les Modèles de Prévision Court-terme de la Production Photovoltaïque.

## *Optimisation of the Use of Multiple Sources of Data in Short-term Photovoltaic Generation Forecasting Models.*


Soutenue par
**Kevin BELLINGUER**
Le 17 Juin 2022


École doctorale n°621
**Ingénierie des Systèmes, Matériaux, Mécanique, et Energétique**

Spécialité
**Energétique et Génie des Procédés**

Composition du jury :

| | |
|---|---|
| Joakim WIDEN<br>Professor, University of Uppsala | *Président*<br>*Rapporteur* |
| Philippe LAURET<br>Professeur, Univ. de la Réunion | *Rapporteur* |
| Ricardo BESSA<br>Senior Researcher, INESC TEC | *Examinateur* |
| Jethro BROWELL<br>Senior Lecturer, University of Glasgow | *Examinateur* |
| George GEORGHIOU<br>Professor, University of Cyprus | *Examinateur* |
| Annette HAMMER<br>Senior Researcher, DLR, University of Oldenburg | *Examinatrice* |
| Georges KARINIOTAKIS<br>Directeur de recherche, MINES ParisTech | *Directeur de thèse* |
| Robin GIRARD<br>Chargé de recherche, MINES ParisTech | *Co-directeur de thèse* |

MINES ParisTech | PSL

*"Cette histoire de développement durable, c'est de la connerie, on est déjà foutu; c'est comme si on exigeait à un cancéreux en phase terminale d'arrêter de fumer sur son lit de mort."*

---

Fight Club

# Acknowledgements

First, I would like to thank Georges Kariniotakis, Robin Girard, and Guillaume Bontron, my thesis supervisors, for offering me the opportunity to work on a very interesting topic and for their good advice throughout this journey.

I also acknowledge the Compagnie Nationale du Rhône for financing this PhD project and providing photovoltaic production data, the European Centre for Medium Range Weather Forecasts (ECMWF) that supplied numerical weather predictions, Transvalor for satellite-derived maps of ground irradiance, and Meteo France for providing us with satellite-based cloud classifications.

# Remerciements

Après ce long périple ponctué de moments de doutes, de désillusions, de mélancolie, et parfois de félicité, mais surtout de franche camaraderie, voici enfin venu le temps des remerciements. Tout naturellement je souhaiterais remercier Georges, Robin et Guillaume pour m'avoir donné la possibilité de réaliser cette thèse mais surtout d'avoir su m'épauler et me guider tout au long de cette aventure. Je remercie Alexandre, Ana, Olivier et Stan de m'avoir accueilli à la CNR et de m'avoir fait profiter de leurs expertises scientifiques. Je tiens à exprimer ma gratitude aux membres du jury d'avoir accepté de consacrer de leur temps à la lecture de cette thèse et les remercie pour les précieux conseils prodigués. Je suis également sincèrement reconnaissant à toute l'équipe PERSEE : Laurent, Fabrizio, Dennis, Alexis, Sophie, Brigitte, Marie-Jeanne, Andrea et Arnaud pour m'avoir accompagné ces dernières années.

Paradoxalement, la thèse, qui est fondamentalement un travail d'ermite, m'a permis de m'ouvrir aux autres et de rencontrer des gens aux horizons variés et aux qualités multiples et diverses. C'est donc, le sourire aux lèvres que je me remémorerai le bureau R09b et ses occupants : Valentin, ce papa hors pair, Fian et le non moins célèbre bus 100, Stéfano et les souris, et Hongxin, cet Heisenberg Sophipolitain. A Shengfei, Biswarup ou encore Akylas, ces marathoniens qui frappent le mur, soyez braves la fin est proche. A ces nouvelles recrues, Owen, Quentin, Luca, Kosta, Yun ou encore Flavien, notre chérubin, vous les ingénus bercés de douces espérances, sachez que la route est encore longue mais que l'erreur est humaine. A Sylvain et Paul, ces compagnons ayant foi en un idéal capillaire supérieur, que j'aurais aimé côtoyer plus souvent. A Wassim pour les courses matinales le long de la côte et à ce whisky qu'il me tarde de déguster ensemble. Jad, j'espère que tes ambitions se réaliseront et qu'un jour j'aurais l'occasion de te rendre visite au Liban. En couchant ces quelques mots sur le papier, je me remémore ces sages à l'allure vénérable, Pedro, Simon, Antoine, Thomas, qui ont su, à leur façon, m'insuffler le goût de la recherche ou de la passion ornithologique, et à Maxime, cet homme à l'œil aguerri parti bien trop tôt. Parmi cette effusion de testostérone quotidienne, un peu de douceur féminine est plus que bienvenue : Riri qui chaque jour se démène pour se créer de nouveaux problèmes et me mettre des paillettes dans les yeux, Anaëlle, cette femme aux milles passions et à l'emploi du temps surchargé, Julia notre *Persesse* au flegme inaltérable, Sacha qui a le don de me (re)donner le sourire et d'égayer ma journée, Kawther et Miaomiao que j'aurais aimé connaître davantage, et enfin Emilie,

cette super maman. En dépit de ces longues heures de dur labeur, je retiendrais avant tout de mon passage dans le sud, ces non moins longues discussions philosophiques, ces flâneries dans la pinède et au-delà, ces excursions au bassin aux poissons, et ces soirées à la plage ou à déambuler dans Antibes.

A tous ces fous qui aspirent à être appelé un jour docteur, je ne saurais trop vous conseiller de vous réfugier dans un exutoire. Pour moi ce fut le Penchak[1]. Bien loin de nourrir une appétence pour le masochisme, ce sport et j'irai même jusqu'à dire cette philosophie m'a permis de me dépasser physiquement et mentalement. Je tiendrais donc à remercier du fond du cœur mon compagnon et ami d'infortune, Adrien, nos mentors Franck, Éric, Antoine, Laurent, Sid, et surtout toute la famille PSDS pour ces bons moments sur le tatami.

Enfin, comment clôturer ces remerciements si ce n'est par ma famille. Il est vrai que quand on (re)vient dans le Nord, on braie deux fois : quand on sort de la gare (sous ce temps maussade et pluvieux, on se languit déjà du sud – histoire vraie...), et quand on repart (le manque des proches se fera ressentir tôt ou tard). Un grand merci donc à Paulette pour sa cuisine chaleureuse, à Mumu pour avoir suivi avec grand intérêt ma soutenance et d'avoir toujours été présente, à Allan pour sa bonne humeur légendaire, et Coco pour son brin de folie et pour qui je m'efforce d'être un modèle au quotidien. J'aimerais consacrer ces derniers mots à Emma pour son soutien indéfectible et qui continue de me supporter malgré les années qui s'égrènent et à m'inonder de sa *mignognitude*. Le temps s'écoule trop vite à tes côtés! Il est de bon aloi de dédier les thèses, je consacre donc cet ouvrage à mes enfants à venir. En conclusion, cette 29ème année aura été des plus riche, et elle n'est pas encore finie...

En fin de compte, il semblerait que les muses ne m'aient pas délaissé.

1. Art martial se caractérisant par une profusion d'ecchymoses et de douleurs variées.

# Preface

Our relationship with energy has changed dramatically since the first industrial revolution which occurred in the late $18^{\text{th}}$ century. Before this turning point, energy consumption was constrained by the limitations of nature (supply of firewood for heating needs, muscle power of men/draught animals and wind/water power for mechanical purposes). Economic growth in such an environment could only be asymptotic. However, this situation changed when societies began to access vast reservoirs of energy built up over the geological era. Humans were no longer dependent on Nature, at least for a while, and exponential growth became possible. From then on, industrial processes developed quickly, in particular thanks to the invention of the steam engine, which converted heat energy into mechanical energy.

During the following two centuries, fossil fuels have been used intensively for around two centuries in a large range of sectors: energy production, industry, transport, chemicals, etc. Since they are finite resources, fossil fuel reserves will eventually run out. Based on this observation, as far back as 1956, Hubbert stated that oil extraction follows a bell-shaped curve: the production rate increases due to an abundance of easy-to-extract resources, then reaches a peak before declining. The debate is still open as to when this peak will be reached, but some energy scenarios expect an oil peak in the next three decades [1]. In 2005, the Hirsch report [2] requested by the US Department of Energy regarding the impacts of peaking oil production concludes that the decreasing supply of oil will lead to an increase in fuel prices as well as political, economic and social instabilities. Fuel price is expected to increase because of (1) population growth; and (2) less easy-to-extract resources.

Human economic growth has been possible thanks to large stocks of carbon-based energies. The combustion of these fossil fuels releases Greenhouse Gas (GHG) emissions which are responsible for trapping heat in the atmosphere. When solar radiation reaches the Earth's atmosphere, part of it is reflected back into space while the other half is absorbed by the surface, the atmosphere and clouds, which in return radiate infrared heat. This heat is trapped by GHG, which are mainly transparent to incoming solar radiation but more absorbent to infrared radiation. This natural process contributed to the development of life on Earth by providing a warm atmosphere, yet these ever-increasing rates of GHG are making the planet hotter, which jeopardises the species living on it. The impact of global warming on climate change is tremendous: glaciers are melting, sea levels are rising, extreme weather events are becoming more frequent (e.g. floods, hotter heat waves, droughts, more powerful

hurricanes, wildfires) [3].

These disasters directly threaten wildlife by disrupting habitats (climate changes alter temperatures and water availability, which in return modify habitat and food availability and force species to adjust rapidly). The temperature-induced phenomenon of coral bleaching is one illustration of how species react to climate change. Nevertheless, through genetic adaptation and a large reduction in carbon dioxide emissions could lead to a 20% to 80% reduction in the bleaching rate of reefs expected by the year 2100 [4].

Human communities will not be spared by global warming. Among the direct and indirect effects of climate change, rising sea level and extreme events such as floods are likely to cause population exodus while drought events and fresh water supply issues will put stress on food-producing systems. Such consequences could jeopardise the geopolitical and economic stability of regions under pressure. Health issues are the object of serious concerns [5]. The risk of illness and death is likely to increase, especially for older groups of people as a result of increased heatwave intensity and frequency. Some studies have highlighted the sensitivity of vector-borne and water-borne infectious diseases to climate: for instance, higher temperature will expand the geographical distribution of malaria [6]. To conclude this paragraph on health issues, it is worth mentioning that the air pollution situation described by Dickens in *Bleak House* is still present today, and is even more worrying. In Europe, around $400,000$ premature deaths per year are due to exposure to PM25 particles [2] resulting from anthropogenic activities [7]. The energy production sector is the main contributor of GHG emissions (25% of worldwide emissions result from electricity and heat generation [8]).

To have a significant impact on GHG reduction, radical measures must be taken. Several options are being investigated and developed around the world. These include reducing our energy consumption through the adoption of sober behaviours (i.e. promotion of the energy sobriety concept which advocates reducing or avoiding energy consumption), and developing energy-efficient building renovation strategies, among others. In parallel, an electricity production paradigm shift has been initiated to replace carbon-based sources with renewable-based sources (e.g. hydro-power, wind and solar power). Nonetheless, these production means are highly weather-dependent, so that the large-scale integration of renewables raises concerns regarding grid stability [3]. Production forecasting appears as a promising solution to deal with the variable and uncertain features of Renewable Energy Sources (RES). Yet, several gaps remain to be filled to enable an adapted use in an industrial environment.

This situation provides the starting point for the present thesis. This work can be viewed as a bridge between the academic and industrial fields, since it mobilises scientific knowledge to respond to operational and tangible issues. This research work was performed at the PERSEE Center [4] at Mines Paristech in collaboration with the Compagnie

2. Particulate matter with a diameter less than or equal to 2.5 $\mu$m.
3. At all times, electricity production must balance consumption.
4. Centre for processes, renewable energies and energy systems.

Nationale du Rhône (CNR), which is France's leading producer of exclusively renewable energy. This work was developed under the supervision of Robin GIRARD (co-director), Guillaume BONTRON (supervisor) and Georges KARINIOTAKIS (director).

# Table of Contents

# List of Figures

**3 Physics-based Modelling**



**4 Data Characterisation**

# List of Tables

# Glossary

**AbM** Analog-based Method

**AC** Alternative Current

**ACF** Auto-Correlation Function

**ADF** Augmented Dickey-Fuller

**AI** Artificial Intelligence

**AM** Air Mass

**AnEn** Analog Ensemble

**ANN** Artificial Neural Networks

**APE** Absolute Percentage Error

**ApEn** Approximate Entropy

**AR** Auto-Regressive

**ARC** Anti-Reflective Coating

**ARIMA** Auto Regressive Integrated Moving Average

**ARMA** Auto Regressive Moving Average

**ARX** Auto-Regressive with eXternal inputs

**ASI** All-Sky Imagers

**BESS** Battery Energy Storage Systems

**BHI** Beam Horizontal Irradiance

**BRL** Boland–Ridley–Lauret

**BRP** Balance Responsible Parties

**BTI** Beam Tilt Irradiance

**c-Si** Crystalline Silicon

**CAR** Conditioned Auto-Regressive

**CARDS** Coupled Auto Regressive Dynamical System

**CCF** Cross-Correlation Function

**CDF** Cumulative Distribution Function

**CE** Cloud Enhancement

**CMV** Cloud Motion Vector

**CNN** Convolutional Neural Networks

**CNR** Compagnie Nationale du Rhône

**COP** Conference of the Parties

**CRPS** Continuous Ranked Probability Score

**CSI** Clear-Sky Index

**CSP** Concentrated Solar Power

**DBSCAN** Density-Based Spatial Clustering of Applications with Noise

**DC** Direct Current

**DCCA** Detrended Cross-Correlation Analysis

**DHI** Diffuse Horizontal Irradiance

**DL** Deep Learning

**DM** Diebold-Mariano

**DNN** Deep Neural Networks

**DST** Daylight Saving Time

**DTC** Distribution Transformer Controller

**ECMWF** European Centre for Medium-Range Weather Forecasts

**EDF** Electricité de France

**EDF** Empirical Distribution Function

**EMOS** Ensemble Model Output Statistics

**ENTSO-E** European Network of Transmission System Operators for Electricity

**ESRA** European Solar Radiation Atlas

**EV** Electric Vehicles

**FAT** Full Activation Time

**FiT** Feed-in Tariff

**FRI** Feature Relative Importance

**FRR** Frequency Restoration Reserves

**GAN** Generative Adversarial Networks

**GBRT** Gradient Boosted Regression Trees

**GHG** Greenhouse Gas

**GHI** Global Horizontal Irradiance

**GMM** Gaussian Mixture Model

**GTI** Global Tilt Irradiance

**HDBSCAN** Hierarchical Density-Based Spatial Clustering of Applications with Noise

**IFS** Integrated Forecast System

**IRENA** International Renewable Energy Agency

**IV** Intraday Variability

**KDE** Kernel Density Estimation

**kNN** k-Nearest Neighbours

**KPSS** Kwiatkowski–Phillips–Schmidt–Shin

**KS** Kolmogorov–Smirnov

**LASSO** Least Absolute Shrinkage and Selection Operator

**LSF** Level Set Forecaster

**LSTM** Long Short-Term Memory

**LT** Linke Turbidity

**LTECV** Loi de Transition Energétique pour la Croissance Verte

**MACC** Monitoring Atmosphere Composition and Climate

**MAE** Mean Absolute Error

**MAEP** Multi Annual Energy Plan

**MAR** Missing at Random

**MBE** Mean Bias Error

**MCAR** Missing Completely at Random

**MDI** Mean Decrease Impurity

**MI** Mutual Information

**MISO** Midcontinent Independent System Operator

**ML** Machine Learning

**MNAR** Missing Not at Random

**MOS** Model Output Statistics

**MPP** Maximum Power Point

**MPPT** Maximal Power Point Tracking

**MRFS** Maximal Relevance Feature Selection

**mRMR** minimal-Redundancy-Maximal-Relevance

**MSE** Mean Square Error

**NA** Not Available

**NLCS** National Low Carbon Strategy

**nMAE** normalised Mean Absolute Error

**nMBE** normalised Mean Bias Error

**nRMSE** normalised Root Mean Square Error

**NWPs** Numerical Weather Predictions

**OPEC** Organisation of the Petroleum Exporting Countries

**P2G** Power-to-Gas

**PAC** Partial Auto-Correlations

**PACF** Partial Auto-Correlation Function

**PAM** Partitioning Around Medoids

**PCA** Principal Component Analysis

**PCs** Principal Components

**PDF** Probability Density Function

**PEM** Proton Exchange Membrane

**PI** Prediction Interval

**POA** Plane-of-Array

**PV** Photovoltaic

**PVPF** Photovoltaic Production Forecasting

**QR** Quantile Regression

**QRF** Quantile Random Forest

**ReLU** Rectified Linear Unit

**RES** Renewable Energy Sources

**RF** Random Forest

**RLR** Robust Linear Regression

**RLS** Recursive Least Squares

**RMSE** Root Mean Square Error

**RNN** Recurrent Neural Networks

**RR** Replacement Reserve

**RTE** Réseau de Transport d'Electricité

**RTI** Reflected Tilt Irradiance

**SampEn** Sample Entropy

**SbPF** Satellite-based Perfect Forecast

**SDSI** Satellite Derived Surface Irradiance

**SDU** Spatially Distributed Units

**SEVIRI** Spinning Enhanced Visible and InfraRed Imager

**SR** Spectral Response

**SSRD** Surface Solar Radiation Downwards

**ST** Spatio-temporal

**STC** Standard Test Conditions

**SVD** Singular-Value Decomposition

**SVM** Support Vector Machines

**SZA** Solar Zenith Angle

**T2M** 2-m Temperature

**TCC** Total Cloud Cover

**TERRE** Trans European Replacement Reserve Exchange

**TL** Transfer Learning

**TSO** Transmission System Operator

**TVAR** Time Varying AutoRegressive

**UCPTE** Union for the Coordination of the Production and Transmission of Electricity

**UNFCCC** United Nations Framework Convention on Climate Change

**VAR** Vector Auto-Regressive

**VARX** Vector Auto-Regressive with eXogenous inputs

**WCWF** Working Condition Without Failures

**WHCO** Weather-Conditioned

**WT** Wavelet Transform

# Nomenclature

**Analogy**

| | |
|---|---|
| $T$ | Target situation |
| $C$ | Candidate situation |
| $D$ | Distance metric |
| $Z_{t+h}$ | Vector of state features at time $t+h$ |
| $\mathcal{T}_M$ | Subset of $M$ closest analog observations |

**Angles**

| | |
|---|---|
| $\theta^S(t)$ | Solar zenith angle formed by the direction of the sun and the local vertical |
| $\alpha^S(t)$ | Solar azimuth angle |
| $\gamma^S(t)$ | Sun elevation angle (i.e. $\theta^S(t) + \gamma^S(t) = 90°$) |
| $\theta(t)$ | Incident angle comprised between the normal to plane and the solar rays |

**Functions**

| | |
|---|---|
| $f_{root}$ | Regression model employed for the mapping of $X_t$ to $y_{t+h}$ |
| $f_j$ | j$^{\text{th}}$ regression tree |

**Operators**

| | |
|---|---|
| $\widehat{.}$ | Expected quantity |
| $\bar{.}$ | Stationarised quantity |
| $.^\intercal$ | Transpose operator |

**Panel characteristics**

| | |
|---|---|
| $\beta$ | Inclination angle of the panel |
| $\alpha$ | Azimuth angle of the panel |
| $L$ | Panel length |
| $s$ | Module row interspacing distance |
| $\tau_{ARC}$ | Transmittance through the anti-reflective coating |
| $\tau_{glass}$ | Transmittance through the glass |
| $\tau_{cover}$ | Transmittance through the cover |
| $\gamma$ | Maximum power correction factor for temperature |
| $A$ | Active cell area of the module |

**Parameters**

| | |
|---|---|
| $\lambda$ | Hyper-parameter that determines the amount of shrinkage in the LASSO |
| $B$ | Vector of the model's parameters to be estimated |
| $\beta_0$ | Intercept |
| $\beta_i$ | Regression coefficients |
| $N$ | Number of observations (or instances) in the response and explanatory features |
| $P$ | Number of variables in the explanatory features |
| $T$ | Number of regression trees |
| $P_c^x$ | Installed capacity of site $x$ |

$N_d$ Number of observations of the day (except nighttime data)

$N'_d$ Number of observations of the day (with nighttime data)

$n_t$ Day of the year

**Solar variables**

$\psi(t)$ Measure of persistence of global radiation level

$k(t)$ Clearness index

$K(t)$ Daily clearness index

$I^{0h}(t)$ Horizontal extraterrestrial irradiance

$I^{0n}(t)$ Normal extraterrestrial incidence irradiance

$AST(t)$ Apparent solar time

$I_0$ Solar constant (here $I_0 = 1361\ W/m^2$)

$AM(t)$ Relative optical air mass

**Temporal variables**

$t$ Time when the forecast is generated

$h$ Number of time steps in the forecast (i.e. forecasting horizon)

$L$ Order of the AR model (here, $L =$ 120 min)

**Variables**

$Y_{t+h}$ Vector of the response variable at time $t + h$ based on elements available at time $t$

$X_t$ Vector of explanatory features which may contain past production and satellite-derived observations as well as NWPs model outputs

$I_t$ Observed irradiance at time $t$

$I_t^{CS}$ Irradiance under clear-sky conditions at time $t$

$P_t$ PV production at time $t$

$S_t^i$ Satellite derived surface irradiance observed at time $t$ and pixel position $i$

$N^{SDSI}$ Number of regressors selected from satellite derived inputs

$N^{NWP}$ Number of NWPs features

$\mathcal{S}$ Set representing the neighbours included in the ST model

$\epsilon_t$ Error term representing random errors or variability from sources not considered in the forecasting model

# Chapter 1

# Introduction

> Smoke lowering down from chimney-pots, making a soft black drizzle, with flakes of soot in it as big as full-grown snow-flakes — gone into mourning, one might imagine, for the death of the sun.
>
> Charles Dickens - Bleak House

## Contents

## 1.1 Context

### 1.1.1 The goals of energy transition

To address the environmental crisis that we are facing, urgent actions must be undertaken to decrease fossil fuel-based emissions and prepare the way for a decarbonised economy.

In 2015, the 21$^{st}$ Conference of the Parties (COP) of the United Nations Framework Convention on Climate Change (UNFCCC) was held in Paris and gathered a total of 196 states. The Paris Agreement established the foundations of countries' energy development to limit the increase in average temperature to below 2°C in this century (with respect to pre-industrial levels). By ratifying this agreement, states committed to propose plans to reduce emissions and to turn to climate-resilient development. Despite, praise for the Paris Agreement, no enforcement mechanism has been defined to assure real commitments. Thus, each country can freely define its objectives. In France, several laws have been enacted, for instance:

- The Energy Transition Law for Green Growth (Loi de Transition Energétique pour la Croissance Verte (LTECV)) (2015) aims to reduce Greenhouse Gas (GHG) emissions by 40% from 1990 to 2030 and to increase the share of renewable energies to 32% of gross final energy consumption in 2030 [9].
- The *Energie Climat* law (2019) aims to achieve carbon neutrality by 2050 with a plan to stop coal-based electricity production by 2022 [10].

To reach these goals, France has set up planning strategies including the National Low Carbon Strategy (NLCS) [11] and the Multi Annual Energy Plan (MAEP) [12].

### 1.1.2 French energy mix

France has a very low-carbon electricity mix thanks to its high share of nuclear power which accounted for around 67.1% of total production in 2020 [13], compared to 23.4% of Renewable Energy Sources (RES) (Figure 1.1).

This situation is substantially due to the *Messmer Plan* initiated after the 1973 oil crisis whose aim was to promote the country's energy independence. This led to the construction of 56 reactors over the next 15 years. Confronted with ageing fleets, in 2014, Electricité de France (EDF) initiated the *Grand Carénage* investment program, which aims at enhancing reactor safety and extending their operation beyond 40 years. Nevertheless, France expressed its willingness to reduce the share of nuclear power through the LTECV, which initially targeted to reduce the nuclear share to 50% by 2025, then postponed to 2035 in 2019.

To promote the installation of RES power plants, governments have developed supportive mechanisms, such as the Feed-in Tariff (FiT), which offers long-term contracts to producers and remunerates them with a cost-based price for electricity injected into the electrical

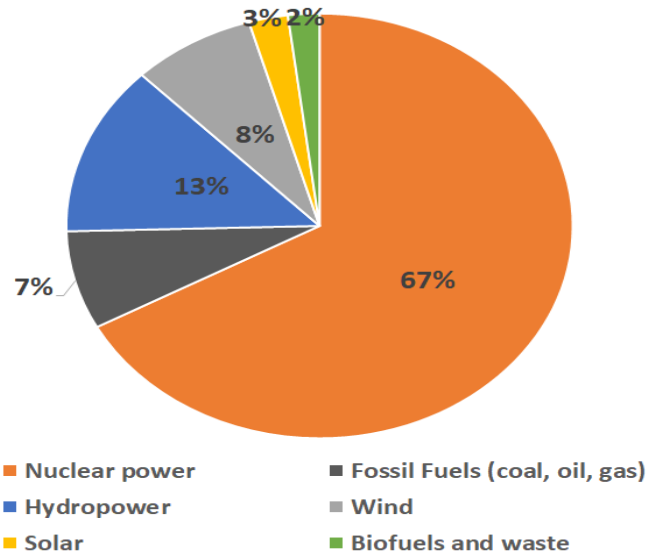



Figure 1.1 – French electricity generation mix in 2020 [13].

grid. In several European countries the FiT schemes have been progressively replaced by a tender support scheme. This compensation mechanism consists in providing producers with a premium tariff in addition to the sale price obtained on the electricity market to cover installation costs and ensure their profitability. This investment security scheme has permitted the growth of RES installed capacity in France (Figure 1.2).

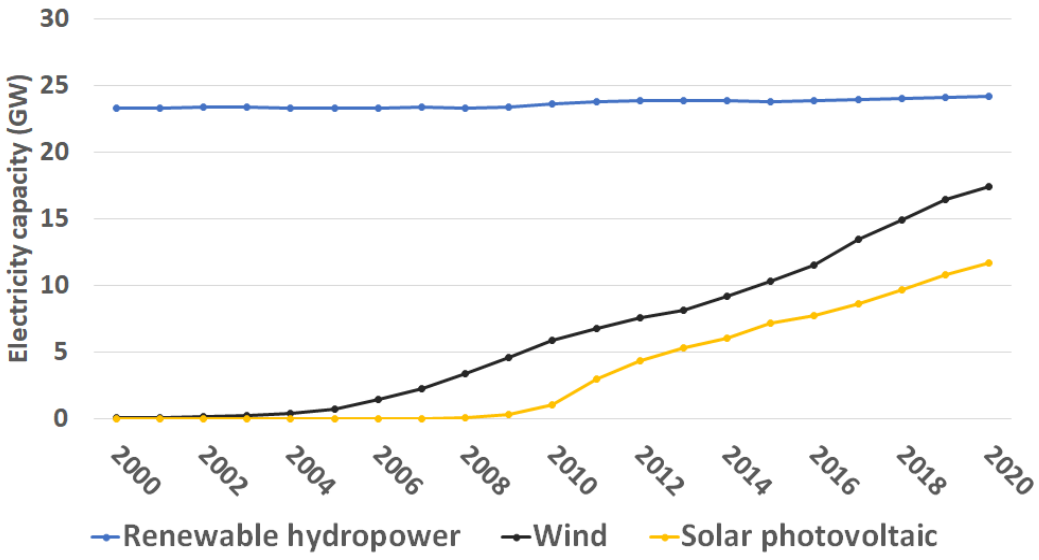



Figure 1.2 – Evolution of installed capacity of RES in France. Source: International Renewable Energy Agency (IRENA) data query tool [14].

This trend is set to continue and even speed up in the future to meet targets. Indeed, the MAEP expects Photovoltaic (PV) installed capacity (roof- and ground-based units) to reach 20.1 GW and 35.1 – 44.0 GW by 2023 and 2028 respectively, while wind power installed capacity is set to reach 24.1 GW by 2023 and between 33.2 GW and 34.7 GW by 2028. Obviously, PV generation will be a key component of the French energy mix, all the more

so as this technology is likely to benefit from an expected further decrease in installation costs: 4% per year for ground-based units and from 5% to 7% for rooftop installations according to the French Ministry of Ecological and Solidarity Transition [12].

## 1.2 The electrical grid

To understand how the electrical grid operates and how it is influenced by RES integration, it is necessary to adopt an holistic approach concerning energy generation, transport and consumption.

### 1.2.1 Evolution of the electrical grid

In the late 19$^{th}$ century, the *age of steam* was replaced by the *age of electricity* [15]. During this period, locations with abundant of resources, such as coal and water, were key drivers of industrial expansion and electrification. Thus, in 1882 Aristide Bergès used the driving force of water to supply his paper mill with energy. At that time, the technological breakthroughs initiated by the first industrial revolution made it possible to obtain steel to incorporate in the turbines structure. Bergès then added a dynamo to convert hydropower into electricity: *la Houille Blanche (white coal)*.

At the beginning of the age of electricity, power generation means and industrial processes were located in the same area for cost production efficiency reasons. The elementary laws of electricity stipulate that the same amount of power can be delivered through a cable by doubling its voltage and halving its current, while reducing heat losses owing to the Joule effect. Thus, a high-voltage current could be used to transfer electricity over larger distances. The first transmission of Direct Current (DC) over a large distance (i.e. 57 km) was performed in 1882 between Miesbach and Munich. In 1891, a 175 km Alternative Current (AC) transmission line was erected between Lauffen and Frankfurt.

In Europe, the electrical network [1] currently in place was essentially developed after World War II. Amidst the ashes left by war, Robert Schuman pronounced on 9 May 1950: "The solidarity in production thus established will make it plain that any war between France and Germany becomes not merely unthinkable but materially impossible". This speech expressed the will to build a peaceful Europe based on cooperation and initiated the European Coal and Steel Community. During this period, the western Europe transmission network became more interconnected, with more exchanges taking place between countries and the establishment of a coordinating structure, namely the Union for the Coordination

1. The electrical grid is an interconnected network that ensures electricity delivery from producers to consumers. Upstream, power plants convert primary energy (e.g. combustion of fossil fuels, kinetic energy of wind) into electricity, which is consumed downstream by end users. Between the two, a set of transmission lines (high voltage electric networks) and distribution lines (low voltage electric networks) associated with substations ensures delivery.

of the Production and Transmission of Electricity (UCPTE) [2]. The main objective of this association was to ensure the optimum operation of electric power plants (e.g. surplus production in countries relying on hydropower could be used to balance a shortfall in other countries whose production was based on oil) [16].

The next step was the *Atomic Age*, which provided the required technology to build nuclear power plants [17], while the oil shocks of the 1970s developed the relevant economic foundations for their expansion. The economic consequences of these shocks led to quadrupling of the price of oil by Organisation of the Petroleum Exporting Countries (OPEC) nations, at that time when most electricity generation involved oil-burning plants. In response to this situation, France pursued its ambition to ensure energy independence by commissioning 44 nuclear reactors between 1978 and 1988 [18]. Owing to its high share of nuclear power associated with a low generation cost and a high pressure of peak load regulation, the country strengthened its transnational interconnections.

Today, the electrical grid is undoubtedly one of the largest and most complex machines ever built by mankind. The synchronous grid of continental Europe counts more than $300,000$ km of transmission lines connecting 535 million customers and disposing of around $1,000$ GW of net generation capacity [19].

### 1.2.2 High-precision machinery

Precisely defining the constraints inherent to the grid would involve a tremendous amount of work and also digress from the subject of this thesis. Nevertheless, a few elements are presented below to provide the reader with a glimpse of the system's complexity.

#### 1.2.2.1 Fragile balance

Electricity is often taken for granted in western European countries. Yet an immense effort is performed backstage to ensure the security of supply and the safe use of production means and consumption devices. The fundamental rule of the network is that *at all times, production and consumption must be balanced*. This rule stems from the fact that electricity cannot be stored in large quantities.

In a perfect equilibrium, the European utility frequency is equal to 50 Hz (this value results from a technical compromise between several phenomena, e.g. low-frequency currents are associated with the flickering of incandescent lamps, while high-frequency currents generate mechanical stress on rotating turbines), while its voltage is equal to 230 V. In practice, the frequency of the electrical grid varies around its nominal value, which is 50 Hz: the frequency decreases when the grid is overloaded - the greater the load on the generator, the slower it spins - but increases when it is underloaded (this phenomenon is illustrated in Figure 1.3).

2. UCPTE is one of the predecessors of the European Network of Transmission System Operators for Electricity (ENTSO-E).

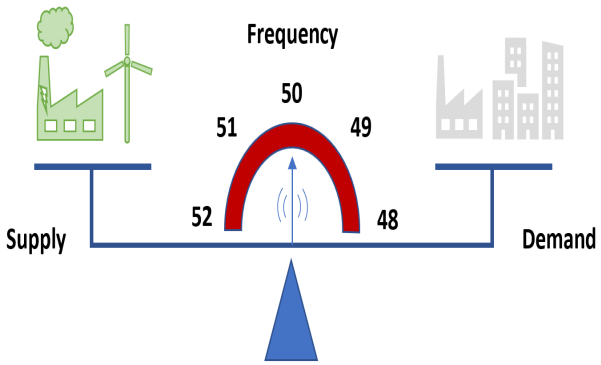



Figure 1.3 – Representation of power balancing

To ensure electricity delivery without compromising the quality of the supply, a set of ancillary services are required. Ancillary services refer to processes or operations used to maintain a balance between production and consumption, to ensure the stability of the transmission, and the quality of the electricity delivered. This kind of services consists in controlling the frequency, or *active power control*, and the voltage, or *reactive power control*. To better understand the actions involved during power balancing, let us focus on frequency control.

#### 1.2.2.2 Frequency control

Frequency deviations continuously occur under nominal operation due to load and generation variations. To prevent the grid from collapsing or damaging generators, the frequency is monitored at all times, and actions are taken to readjust the frequency to its nominal value when needed. This task is generally the responsibility of the Transmission System Operator (TSO). Each country has its own set of technical rules regarding frequency balancing, and so here we focus on the ENTSO-E guidelines. In normal operation, the frequency deviations must be maintained below $\pm 1\%$ of the nominal value [20]. When the frequency of the electrical grid reaches an emergency condition (e.g. due to an incident like the loss of a generator) a frequency control strategy is initiated by the TSO to restore power exchanges in its control area within a maximum of 15-min at the latest. This strategy comprises at least three (partially overlapping) balancing mechanisms with specific features, illustrated in Figure 1.4.

1. The first level is **primary control** (also called R1). This is the fastest response to frequency deviation; it is automatically initiated by production means just a few seconds after the incident and lasts until the offset of deviation. Its Full Activation Time (FAT) (i.e. the period between the activation request and the full delivery of the reserve) is 30 s. When a frequency disturbance event occurs, the frequency variation is softened by the kinetic energy stored by the set of rotating masses within the synchronous grid; this is called the *inertial response*. Then, the turbine speed controllers that are already online, the so-called *spinning reserves*, increase the power

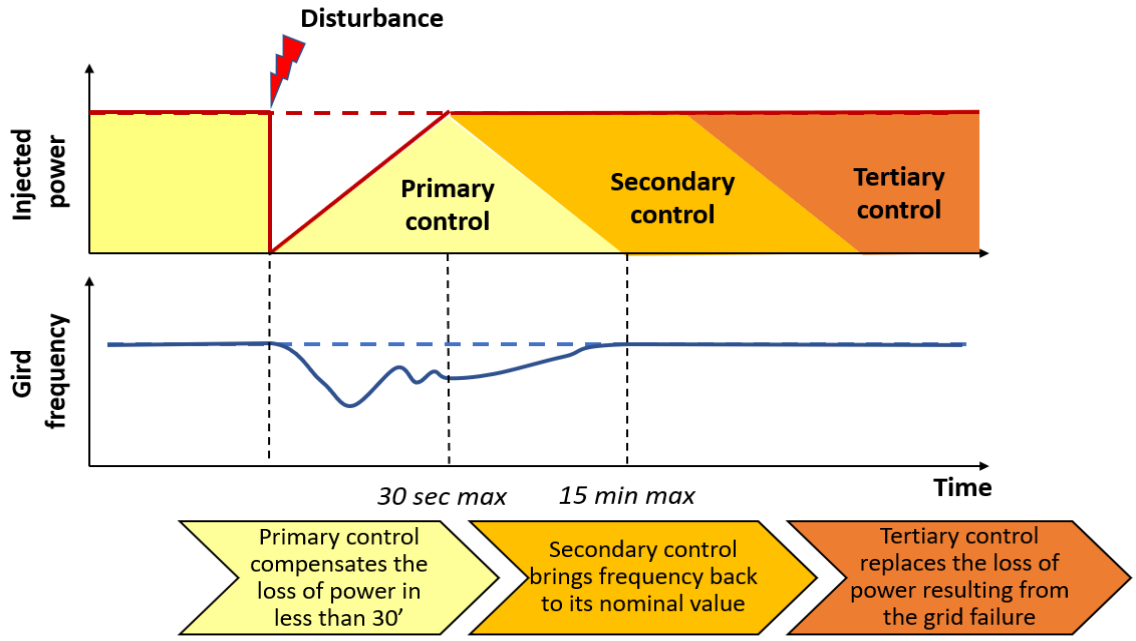



Figure 1.4 – Illustration of the activation of balancing services in France, inspired from [21].

output. This action aims at rapidly stabilising the frequency at a quasi-steady-state value within permissible limits but without restoring the system frequency to its reference value. Dispatchable power plants are typically in charge of primary control; RES can also be involved in this process through curtailment actions (to reduce production) or by intentionally working in clipped operation mode (i.e. the actual power is lower than the nominal production to permit an increase in production if needed). These production adjustments are performed by employing the Maximal Power Point Tracking (MPPT) of inverters, which constantly look for the best operating voltage to get the highest possible power from the arrays. Batteries can also be integrated to mitigate grid fluctuations.

2. The second control action to be automatically committed is **secondary control** (R2). This second level aims at restoring the nominal value of the system frequency, the reserve of each generator used during the primary control, and the scheduled cross-border exchanges with adjacent control areas. This process must be completed within 15-min at the latest. Like primary control, secondary control is composed of the spare capacity at the disposal of the TSO: additional generation capacity can be requested, or some production means can be stopped.

3. **Tertiary control** (R3) is primarily used to release the reserves enrolled during secondary control and return to a state of readiness when an equilibrium is reached. This process is not automatic but is requested manually by the TSO. Since it is the last level to be activated, more sources of flexibility can be considered (e.g. flexibility services provided by industrial processes). In France, tertiary control is composed of 1 GW capacity, which can be activated in 15-min and for 2-hour [22]. It can also

be activated to support the secondary control process in case of large incidents and consequently to free power reserves activated during primary control.

4. In addition, some TSOs maintain a Replacement Reserve (RR) to release previously activated Frequency Restoration Reserves (FRR) in case of new disturbances. This new product is part of the Trans European Replacement Reserve Exchange (TERRE) project, which aims at developing a European central platform for the exchange of balancing resources (both an increase and decrease of active power). In France, this reserve is composed of 0.5 GW which can be activated in less than 30-min and for 1.5-hour [22].

The reserve services are organised into contracts with the TSO and electricity producers and large energy users to provide temporary extra resources or to request a reduction. These services can also be provided through electricity import and export. These balancing products are engaged through auctions within the balancing market, which is a real-time market that ensures the balance of the power system.

### 1.2.3 Towards a smart power system with...

The electrical grid, by linking production areas to load centres, acts as the backbone of the renewable-based energies transition. To cope with the ever-increasing variability of consumption and production, the energy sector is undertaking a series of investments to modernise the electrical grid. Thus, automation and communication devices are being explored for their capacity to measure and eventually monitor the energy flows at any time [23].

#### 1.2.3.1 ... new uses

The shift to a sustainable low-carbon economy is closely tied to the electrification of some high GHG emitting sectors. These include the transport sector and the gas industry. In 2016, the former accounted for around 27% of European GHG emissions, slightly more than the power generation and industry sectors [24]. To overcome this issue, the current vehicle fleet running on fossil fuels is being replaced by greener technologies, such as Electric Vehicles (EV) and hydrogen vehicles. In 2020, more than 10 million light-duty EVs were on the world's roads, this number could reach 145 or 230 million in 2030 depending on the scenario [25]. The coupling of the electrification of such sectors and demographic growth will result in an increase in electricity demand and the rise of new challenges regarding grid constraints: for instance, the simultaneous charging of a large EV fleet might generate grid congestions or peak demands.

#### 1.2.3.2 ... new services

Far from being a burden on the grid, these new technologies can offer valuable opportunities to improve the flexibility and security of the power system, especially in a context of

high RES penetration. Different options are investigated in the literature. When considered in vehicle-to-grid mode, EV can be used as a distributed energy storage unit: energy from vehicles' batteries is injected into the grid to flatten out RES intermittency, while reducing energy costs [26], or to shave demand peaks [27]. On the contrary, EVs are charged during consumption dips, or when electricity pricing is at its lowest. Another option concerns hybrid energy systems by which solar energy can be stored in an energy carrier such as hydrogen [28] or in Battery Energy Storage Systems (BESS). This first approach based on Power-to-Gas (P2G) architecture (i.e. a system that converts electricity into hydrogen via electrolysis), produces hydrogen, which can be directly injected into the gas network or later converted into electricity thanks to a fuel cell. In [29], the authors show that coupling RES-based units with P2G technology can reduce the required capacity and curtailment of plants. Moreover, grid-connected Proton Exchange Membrane (PEM) electrolyzers and fuel cells technologies appear to be a feasible solution to provide the electrical grid with ancillary services in terms of frequency regulation, voltage control and congestion management [30]. It is worth mentioning that most of these solutions, despite being technically feasible, are not commercially viable at present: e.g. [31] highlights that the benefit of installing a BESS is low compared to its installation cost.

#### 1.2.3.3 ... new actors

This intelligent grid is designed to allow more flexibility from small distributed generation units and consumption entities. Today, the electrical grid can still be viewed in a binary way, with the producers on one side and the consumers on the other. Nevertheless, this dichotomy is expected to diminish in the near future owing to the decentralisation of production and the fact that consumers will play an increasing role in stabilising the network. Thus, a new kind of actor is emerging, called the *prosumer*, who produces electricity thanks to renewable-based means (e.g. rooftop PV units) but also consumes electricity from the grid. The interactions between households and utilities will increase in order to better manage network stability. For instance, during a consumption peak, some devices such as water heaters could be automatically shut down to smooth the demand curve.

### 1.2.4 Electricity markets

#### 1.2.4.1 From a monopoly situation to a liberalised market

EDF was founded in 1946 as a result of the nationalisation of a hundred energy producers, TSOs and distributors, and went on to become the main electricity generation and distribution company in France. This state monopoly was abolished in the 1990s by a European directive aimed at the gradual liberalisation of energy trading as well as the unbundling of the main activities (transmission and distribution activities maintain their monopoly but are placed under regulation schemes). This liberalisation aimed at opening the market to new energy providers and establishing a competitive environment supposed to be beneficial

for the customers through a cost of electricity reduction thanks to the competition. To sum up, a shift took place from centralised generation, characterised by significant government influence, to a free, competition-based paradigm open to aggregators. An aggregator is an entity that aggregates a number of disparate consumers and/or producers and acts as a liaison between them and the wholesale markets [32]. In this respect, in addition to being France's leading producer of exclusively renewable energy, Compagnie Nationale du Rhône (CNR), which supported this research project, also acts as an aggregator on behalf of its customers. It manages a portfolio of several RES plants with a perspective of optimising their production sales on the markets.

#### 1.2.4.2 Several markets for several products

Electricity is a tradable, fungible commodity that can be sold or bought on several electricity markets that possess their own properties. Depending on the time horizons, these markets can be classified into four types:

1. **Futures and forwards** markets are designed for purchasing products up to several years ahead to secure business against price fluctuations.
2. **Day-ahead** markets are operated once a day and enable the hourly sale and purchase of products for delivery on the next day. This market is liquid [3].
3. In **intraday** markets, participants trade continuously. These markets offer the possibility to trade electricity up to 5-min before delivery and to adjust balance. The exchanged quantities of electricity are smaller than on the day-ahead markets inasmuch as the purchase results from unplanned consumption, and prices are generally less attractive. The liquidity of this market is lower than that of the day-ahead market.
4. Unlike the two previous markets, which trade energy for the future, **balancing** markets deal with the purchase of services that can be committed to guarantee grid balancing. The real-time balancing of the grid is performed by the TSO, which has previously purchased R1, R2 or R3 balancing products.

#### 1.2.4.3 Financial costs

To ensure secure operation of the grid, the TSOs have to maintain a balance between production and consumption within their control areas. They are supported in this mission by Balance Responsible Parties (BRP), such as CNR, that must ensure adequacy between energy injections and extractions within their balance perimeters. These perimeters are portfolios of energy suppliers and consumers. Any mismatches observed in balance perimeters result in financial penalties incurred by the TSO which had to engage reserve

3. Liquidity refers to the degree to which an asset can be quickly purchased or sold without inducing price variation or significant transaction costs. A good indicator of the market liquidity is the volume of transactions performed without affecting transaction prices: in this case a higher trading volume is associated with a higher level of liquidity [33].

mechanisms to maintain grid stability. Therefore, accurate forecasts are of prime interest for BRPs dealing with RES units to mitigate these imbalance costs.

In addition, to promote large-scale integration of RES into the grid, states have proposed incentive policies based on FiT or subsidies. Such mechanisms are coming to an end, since producers are expected to bear the costs generated by the balancing of their own production. This compels RES producers to integrate energy markets. This process requires producers to have relevant forecasting tools to submit production schedule bids. Small entities that do not possess such resources can outsource forecasting and bidding to aggregators. RES production, which is variable by nature, is likely to deviate from schedules. Depending on the regulatory framework of the country, this may lead to settlements because of a contractual mismatch between the energy sold and the actual energy provided. For instance, the Midcontinent Independent System Operator (MISO) [4] imposes a real-time price for imbalances but without deviation penalties, while on the contrary Spain applies financial penalties if the actual production deviates from the forecast, and in India, penalties are enforced when forecast deviates by more than 30% [34].

#### 1.2.4.4 Value of forecasting

The notion of *value of forecasting* refers to the potential economic benefit resulting from an improvement in the forecast accuracy (i.e. reduction of the error). The financial impact is obvious when it comes to power producers able to minimise their imbalance costs via better forecasting tools. In this respect, [35] highlights that a reduction of day-ahead forecasting uncertainties correlates with a profit increase from computing deviations between the scheduled and actual production of a 1.86 MW utility-scale plant operating on the Iberian electricity market. Whereas, the economic value of forecasting is not restricted to imbalance charges, it may positively impact the entire power system operation by reducing operating costs, while contributing to system reliability and security. In [36], the authors point out that an improvement in solar power forecasts generates a reduction in the commitment of inefficient power plants (e.g. gas and oil turbines) and PV curtailments, which leads to a reduction in the overall operating costs. Moreover, large forecasting errors concerning RES production increase intra-day prices and the size of the system imbalances needed to accommodate RES variability and uncertainty [37].

## 1.3 Towards a sun-powered future?

The theoretical potential of solar energy striking the Earth's surface in the space of one and a half hours represents more energy than global energy consumption in 2001, which makes sunlight the highest *theoretical* potential of RES [38].

4. A North American system operator.

### 1.3.1 Harvesting technologies

Two technologies are typically used at utility scale to produce electricity from sunlight: (1) PV systems and (2) Concentrated Solar Power (CSP) systems. The first is based on the photovoltaic effect which converts light into electricity, while the second uses lenses or mirrors to concentrate sunlight to heat a fluid (typically water), which then drives a steam turbine or directly feeds an industrial process like water desalination [39]. CSP technology has the ability to store energy (in the form of heat), which mitigates production irregularity issues [40]. For domestic applications, solar energy is usually harvested for water heating purposes as well as to produce electricity via rooftop PV panels. Fields of research explore alternative ways of converting solar energy, one may mention the use of artificial photosynthetic systems able to convert solar energy into solar fuels such as hydrogen from water [41]. Thanks to impressive cost reductions in the last ten years, today PV production systems represent the most dynamic market (compared to CSP systems) [42].

### 1.3.2 Photovoltaic effect

The active part of a solar cell is a wafer composed of silicon, which is a semi-conductive material. This wafer can be broken down into three layers:

1. The top layer is a thin layer of silicon doped with group V atoms such as phosphorus. The doping process consists in introducing impurity atoms into the semi-conductor with the aim of improving its conduction characteristics. Phosphorus elements have five electrons in their outer orbitals while silicon elements only have four. This fifth electron has nothing to bond to, and can freely move, which makes this layer more conductive. Due to this electron excess, this layer is called "negative-type" (n-type) as it favours the transport of a negative charge.
2. The bottom layer is composed of silicon doped with group III atoms such as boron. This dopant has three electrons in its outer orbital. A missing electron can be viewed as a positive charge, hence the name of this layer: positive-type (p-type).
3. When the n-type and the p-type are put together, we observe a displacement of charge carriers: (1) the nearest extra electrons on the n-type side migrate towards the holes of the p-type layer and (2) alternatively the extra-holes from the p-type sides migrate towards the atoms on the n-side which need holes.

On the one hand, the junction between each layer becomes depleted of charge carriers (electrons and holes cancel each other out), and loses its conductive properties. On the other hand, in the n-type layer, the region near the junction becomes positively charged (the phosphorus atoms are fixed in the lattice and one electron is missing) while the p-type region near the junction becomes negatively charged. Therefore, a potential difference is formed between the top and bottom layers. An equilibrium is reached when (1) the diffusion process (i.e. the migration of electrons to holes and vice versa) tends to increase

the potential difference and (2) the resulting electrical fields, which tends to oppose the charge carriers' displacement, counteract each other. In such a component, the electrons can only flow through the n-type side to the p-type side: this is known as a diode.

The PV effect occurs when a photon transfers its energy to an electron from the junction area, pulling it out of its atom and leaving a hole. Due to the electric field, the electron is driven to the n-type layer, while the hole moves towards the p-type conductor. Then, two electrodes located at the upper and lower part of the solar cell permit the generation of an electrical current (the electron goes towards the hole).

### 1.3.3 A higher share of photovoltaic in the future

Currently, the penetration level of PV production in France is still modest (i.e. 2.5% for around 10 GW of installed capacity at the end of 2020 [13]). However, this coverage rate could increase up to $9\% - 37\%$ (which represents respectively 40 GW and 185 GW of installed capacity) in 2050 according to the scenarios provided by Réseau de Transport d'Electricité (RTE) [43]. These scenarios, based on the NLCS and the MAEP, consider two main directions for France: either a fully renewable-based production or a mix with new nuclear units.

### 1.3.4 Solar production variability and predictability

RES are characterised by high variability (i.e. a change in generation during a certain period of time), and limited predictability (i.e. due to the chaotic nature of the atmosphere, future generation is difficult to assess accurately and, therefore it is more or less uncertain) (Figure 1.5). This induces a limited forecastability inherent to the modelling strategy involved. At this point, it is worth defining the concept of predictability and forecastability. Both terms are often used interchangeably but differ by their meaning. Authors in [44] propose the following definition: predictability studies how trajectories of the system diverge, while forecastability describes how a model trajectory diverges from the system trajectory.

PV generation variability can be broken down into a deterministic component and a stochastic component. The Sun's motion in the sky is governed by fully understood deterministic astronomical laws, which are responsible for intra-daily and intra-yearly generation variability. Intra-daily variability is characterised by a bell-shaped curve (i.e. low production levels are associated with low solar elevation angles, while peak production is observed when the Sun is at its highest), and seasonal variability is associated with lower production rates during wintertime as well as shorter days (Figure 1.6). In addition, atmospheric conditions such as cloud distribution or air turbidity affect the amount of light reaching the ground as well as its spectral distribution.

Although the physical equations governing the atmospheric states were established in the $19^{\text{th}}$ century, today it is still not possible to provide perfect weather forecasts. This shortcoming is due to the chaotic nature of the atmosphere. One of the most popular

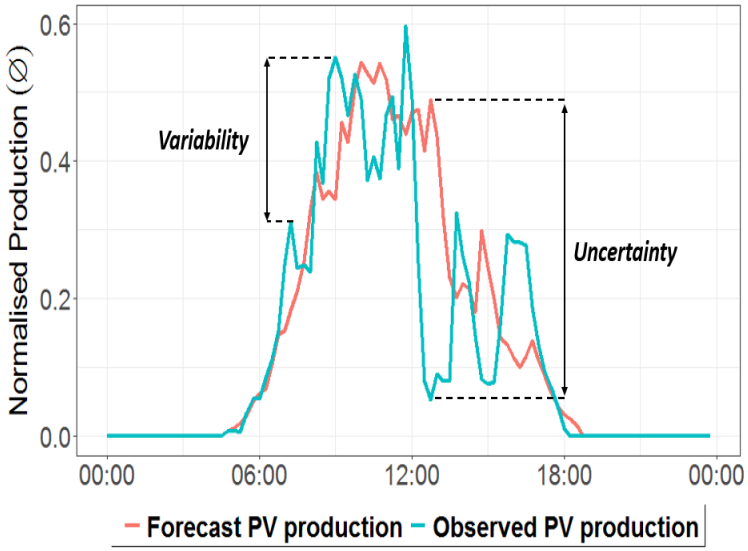



Figure 1.5 – Schematic representation of the concept of variability and uncertainty of PV generation forecasts.

images of chaos is undoubtedly the *butterfly effect*, which states that small changes in the initial state of a deterministic nonlinear system can result in a very different future state. Thus, the deterministic nature of the atmosphere does not make it fully predictable because of the impossibility of observing every detail of its initial state. Small errors in the initial state are amplified with time, which leads to large errors in forecasts. The chaotic nature limits the predictability to about 14 days.

## 1.4 Scope of this thesis

Among the conceivable options to deal with the variability of RES generation, this PhD focuses on the forecasting field. RES forecasting appears as a cost-effective option to anticipate power imbalances and thus to optimise the use of flexibility solutions or traditional adjustment means. Typically, the Photovoltaic Production Forecasting (PVPF) domain can be split into two groups (1) day-ahead forecasts that are used to find an optimal energy trading strategy in the day-ahead energy market, but also for power system scheduling or reserves estimation, and (2) intraday forecasts, which are vital to grid operators to define a balancing schedule (i.e. spinning reserves or demand response that can be engaged to tackle an expected imbalance), to manage congestion, and to define offers for the intraday market. Our partner, CNR, which plays a dual role of aggregator and BRP has to balance its renewable energy portfolio, while optimising profits on energy markets. This work focuses on short-term forecast horizons. The short-term horizon terminology has not been precisely defined by the PVPF community; here it should be understood as horizons ranging from 15-min to 6-hour ahead. The typical period of interest is 15-min, which implies that we need to anticipate production at these horizons.

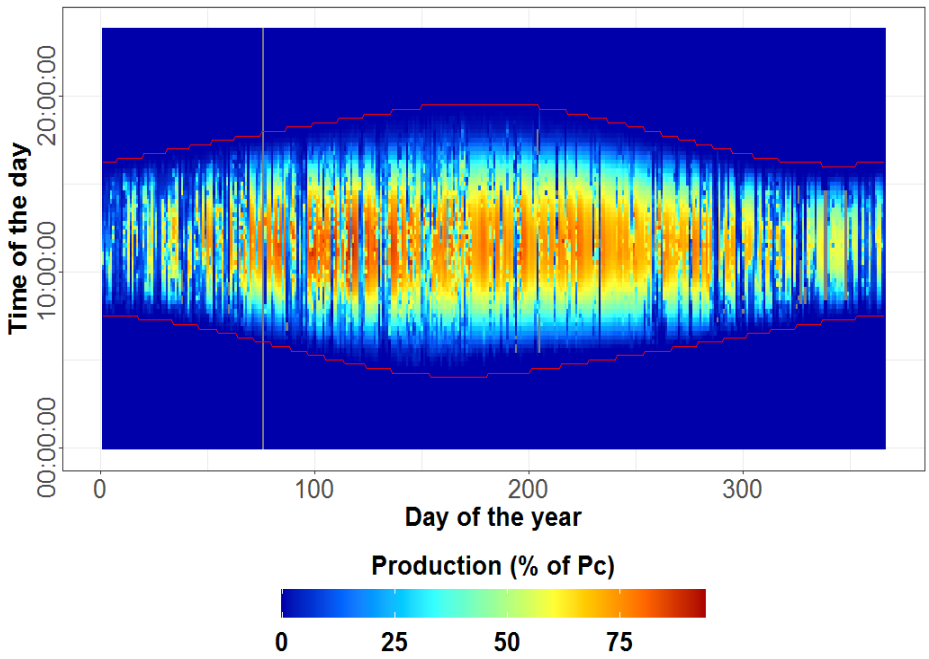



Figure 1.6 – Intra-day and seasonal variability of PV generation. The plot refers to measurements from a single PV plant that is normalised by the installed power of the plant. Red lines represent sunrise and sunset times.

Considering the analysis of the state-of-the-art, which is developed throughout this document, in this work we fix as objectives to propose a forecasting approach that meets the following requirements: (1) that is as simple as possible to make it easy its use by a wide range of end-users (from academics to RES forecasters and PV plant operators), (2) that is suitable for online application (i.e. robust and rapid fitting), and (3) which is extensible in the sense that additional inputs can be included and forecast horizons can extended.

As pointed out by Tawn in [45] future research directions suggest a greater prevalence of probabilistic forecasting. Nevertheless, we chose to focus on point forecasting before turning to probabilistic modelling. Our pursual of this counter-trend, may be justified by several arguments. The first is that today in the state-of-the-art there is a tendency to go to complex modelling approaches that are often black-box types of models. In this work we made the choice to go back to the basic directions following a more *physics informed* approach in the design of our prediction models. This is because we want to have a better interpretability of the results. Second, deterministic forecasting provides us with the possibility to express and analyse the relationships between the response and explanatory features. Moreover, point forecasts are easier to analyse inasmuch as probabilistic forecasts also include reliability properties to assess. Last, to some extent, we may assume that for the timescale under study (namely from a few minutes to 6-hour ahead), the chaotic nature of the atmosphere does not prevail. This premise may not hold true for in tropical climates.

## 1.5 State-of-the-art and positioning

In this Section we present an introductory analysis of the state-of-the-art that aims at revealing the main gaps observed in the short-term PVPF field. For the sake of completeness, additional information regarding the state-of-the-art will be supplied throughout the following chapters.

RES generation forecasting clearly plays an important role to meet the challenges of high shares of RES integration in power systems and electricity markets In this regard, over the last decade, PVPF has been a very active field of research as reflected by the increasing number of related publications (Figure 1.7), and significant progress has been made. Interested readers may refer to [45–48], which provide fairly complete literature reviews. In 2016, [49] stipulated, that PVPF was still an immature domain on the grounds of the late development of solar power penetration -in comparison with wind power forecasting, which dates back to the 1980s [50]. In light of the number of publications and international collaborations dealing with renewable forecasting (e.g. the European Horizon 2020 project Smart4RES [5] [51]), it is plausible to claim that this assessment is still pertinent.

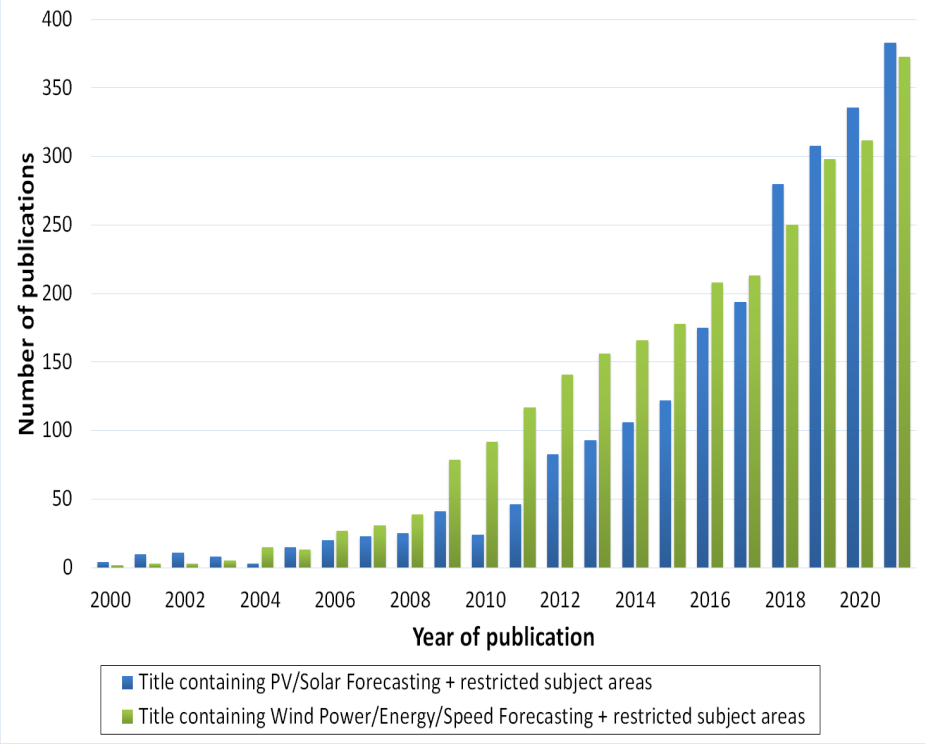



Figure 1.7 – Evolution of the number of publications related to wind and PV generation forecasting fields in the Scopus database.

To improve forecasting performances, three options are conceivable: (1) improve the accuracy of the forecasting models, (2) consider high-quality sources of information, or (3) a combination of both.

5. http://www.smart4res.eu

### 1.5.1 Model classification

Traditionally, PVPF models can be divided into three groups with respect to the modelling process: physical models, statistical models, and hybrid models that result from a combination of the two.

Physics-based models are parametric models that consider PV generation as a *white-box* where the irradiance-to-electricity conversion process is modelled explicitly via analytical expressions. In this paradigm, the modelling of the atmosphere effects over incoming irradiance is done by Numerical Weather Predictions (NWPs) models. To be efficient, this approach requires accurate irradiance forecasts and thorough knowledge of the physical characteristics of the conversion process at stake. Physics-based forecasting models are particularly relevant for recently built PV systems, which do not possess past production records to train statistical models.

On the other hand, statistical approaches do not presume any knowledge about the physical process. In contrast to the physical approach, this is a data-driven one, which infers conversion laws based on historical time series. This needs large datasets to train models based on statistics or Machine Learning (ML) algorithms. Moreover, statistical models can consider inputs' systematic errors (e.g. this kind of models is able to integrate measurement errors of inputs). Based on the previous analogy, this approach can be identified as a *grey-box* or *black-box* modelling according to the degree of abstraction/transparency of the models. With the increasing popularity of complex approaches such as Deep Learning (DL) [6], the interpretability of ML-based models is receiving increasing attention in various fields. The notion of interpretability can be defined as the ability of a human user to understand the links created by the model between inputs and outputs. In other words, an algorithm is interpretable when a human can understand how it (the algorithm) is using the input information to generate the output. This notion is closely linked with the model's complexity: the more parameters a model has, the more difficult its analysis will be. Thus, when it comes to analysing DL models with several hidden layers such understanding is challenged: hence the *black-box* terminology. This opacity becomes a burden when it is necessary to identify which patterns in the data are the most relevant for predictions. It is worth mentioning that other ML algorithms such as Random Forest (RF) have specific tools or "embedded mechanisms" to determine features importance.

Statistical approaches are the most frequent in the literature [47] and commonly outperform physical modelling [52]. In the author's opinion, the literature dedicated to statistical-based modelling principally focuses on the development of advanced forecasting models, to such an extent that the physical properties of the parks are often left out (i.e. models are supposed to learn and mimic the conversion process on their own). Thus, it is quite common to provide regression models with raw NWPs model outputs, such as Global Horizontal Irra-

6. DL is a branch of ML based on Artificial Neural Networks (ANN). The adjective "deep" refers to the use of multiple layers in the network to extract high-level features from inputs.

diance (GHI). On the contrary, Physics-based models explicitly model the conversion of GHI into electricity power taking into account the influence of temperature and the orientation of panels for instance.

?

**Main Research Gaps - *Coupling of physics and statistics***

Therefore, physics- and statistics-based modelling are two opposed ways of estimating electricity production from irradiance forecasts. These two fields seldom merge. Still, it could conceivably be possible to improve forecast accuracy by including system-based knowledge within statistical models. Such an approach could be implemented as a preprocessing step that converts forecast irradiance into predicted electricity power. Then, the statistical model would polish the output (e.g. by dealing with systematic errors of NWPs model outputs). In addition, this could be seen as an attempt to preserve model interpretability by injecting physics-based knowledge. The coupling of physics and statistics raises two main issues:

1. Can statistical modelling strategies benefit from the integration of system-based knowledge?
2. With the advent of ML- and DL-based models and an observed tendency to black-box modelling, can we preserve models' interpretability without compromising performances?

### 1.5.2 Growing diversity of information sources

Several sources of information are currently investigated in the PVPF-related literature. Each one possesses different characteristics, which make them horizon-specific (Figure 1.8).

For short-term PVPF (i.e. from a few minutes to 6-hour ahead), endogenous inputs, namely past PV production measurements, are typically the main drivers. For this horizon range, statistical models fed with lagged observations are able to assess the production dynamic of the system. PV production parks are made up of several sub-components that drive the conversion process (e.g. inverters, transformers) of DC produced by PV strings, namely series-connected sets of modules.

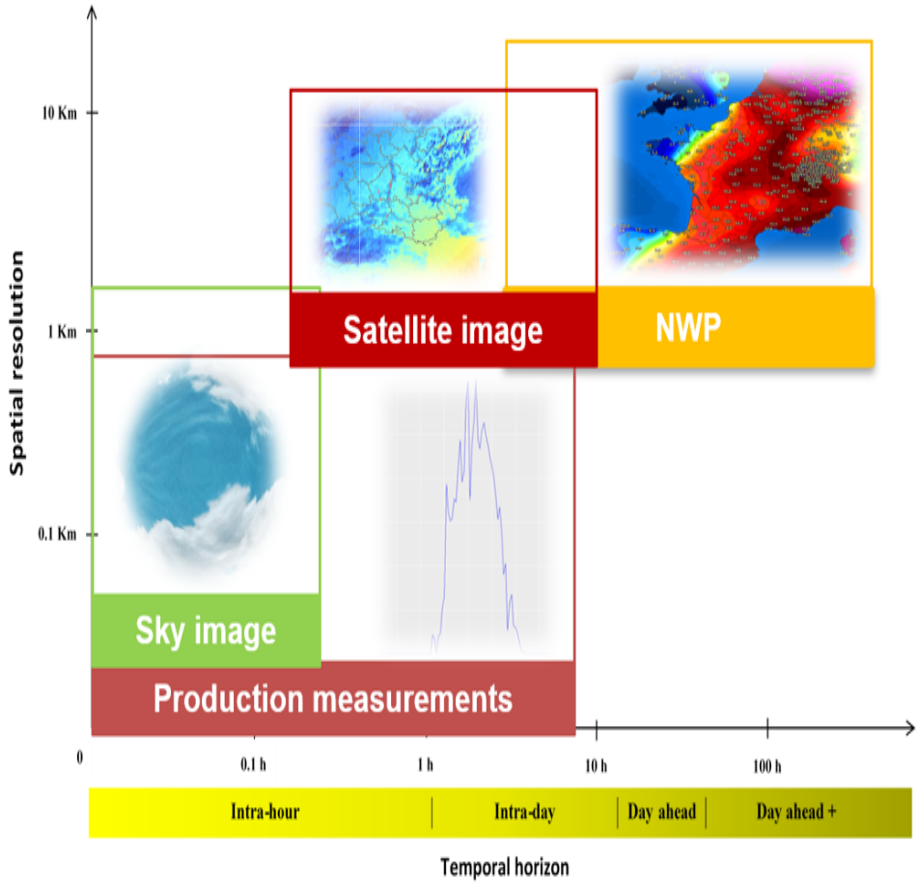



Figure 1.8 – Classification of inputs type according to the forecast horizons and the spatial resolution (inspired from [46, 47]).

?

**Main Research Gaps - *Quality of production observations***

The different elements of this chain are subject to malfunction, which degrades the production rate of the whole plant. These failures require special attention from forecasters as they may negatively impact the forecast accuracy of the statistical model trained on these faulty observations. Yet in the PVPF-related literature, this topic receives little coverage: most of the time researchers content themselves with elementary verification rules. As a result, it is legitimate to delve into the quality analysis of production observations. This raises the following questions:

1. What resources can be used to assess observations associated with production failures?
2. What is the impact of malfunction behaviour over forecast accuracy?

For shorter lead times (i.e. *nowcasting*), All-Sky Imagers (ASI) are typically used. They

offer real-time, on-site observations of passing clouds by taking photographs of the sky dome above their point of installation. A network of several cameras (such as the Eye2Sky ASI network [53]) makes it possible to estimate the spatial distribution and height of cloud structures, and to compute cloud shadow maps [54]. A series of images are then used with dedicated tools (e.g. optical flow or block-matching algorithms) to predict the cloud motion and the corresponding shadow projection.

Observations of cloud cover can also be taken from above the sky. Space-borne photographs of Earth provide valuable information regarding the cloud distribution and optical properties at a higher scale than what can be achieved with ASI. This thus extends the forecast horizon: in astrophysics the observation *the further out we look in space, the farther back in time we see* holds, while in the present context, the correct statement would be *the further out we look in space, the farther into the future we see*[7]. However, this extension is performed at the cost of lower spatial resolution. The latter depends on the satellite technology and the pixel position in the image, for instance, Meteosat Second Generation satellite generates images with resolutions ranging from 3 km at the nadir to more than 12 km on the edges of the planet. They are updated regularly (e.g. every 15-min) practically without delay of delivery and are typically used in the literature for prediction horizons ranging from a few minutes to 6-hour ahead (e.g. [55]).

Satellite images provide spatially distributed observations of the atmosphere. Without resorting to such complex devices, it is possible to employ off-site sensors like local weather stations to get a glimpse of the weather situation in the vicinity of the power plant [56]. Such approaches rely on the Spatio-temporal (ST) dependencies that may exist between the sensors and the site of interest. Their accuracy is highly dependent on the density of sensors and their spatial distribution. Similarly, one can consider the production measurements of a nearby PV system as a proxy of the level of irradiance [57, 58]. This idea seems to have first been applied in the PVPF field in 2011 [59]. Such an approach is appealing for a producer/aggregator inasmuch as it does not require extra sources of information or additional costs.

For higher lead times, such as day-ahead, NWPs become the main source of information. They are issued by physics-based numerical models that resolve the governing equations of the atmosphere[8]. Such models are greedy in terms of computational resources, which compels to reduce output precision. Typically, NWPs have an hourly temporal resolution and a grid resolution of around 10 km. Such features provide relevant information regarding atmosphere trends but fail to capture the weather variability at the site. In this respect, it is worth mentioning that ongoing studies are working on the development of RES-dedicated weather forecasts with $10-15\%$ improvement using various sources of data and very high-

7. The future temporal horizon being limited by the cloud's lifespan.

8. The governing equations are resolved at the discretization scale, while lower scale phenomena are parameterized (i.e. they are approximated because the processes involved are too small/brief or complex to be explicitly integrated, and to reduce computational costs).

resolution approaches [60]. That is why, for the short lead-times considered in this paper, NWPs are often neglected in the literature to the advantage of ST data. Nevertheless, we will consider them as a potential input to be analysed since several results [55, 61] indicate that they contribute to the improvement of forecast accuracy.

Therefore, each source of information possesses its own temporal and spatial characteristics, which allows them to assess specific weather phenomena. The natural reaction is to combine all these inputs within forecasting models. In [55], the authors show that combining ground observations with exogenous data (satellite-based observations and NWPs) improves the accuracy of intra-day irradiance forecasts.

### 1.5.3 A need for model adaptivity

PV generation depends on a large number of meteorological variables such as irradiance, cloud cover, airflow motion, ambient temperature and even air humidity. The combinations and interactions of these variables lead to a large range of weather states associated with significant varied dynamics. For this reason, NWPs provide valuable information to PVPF models by giving them information on the expected atmospheric state and how it is likely to influence PV production.

The predicted weather information can be integrated in the PVPF modelling chain in two different ways: either explicitly (i.e. additional explanatory features) and/or implicitly (i.e. as state variables), which permits local modelling.

#### 1.5.3.1 Explicit integration

The most straightforward method considers NWPs as additional explanatory features inside the PVPF model (i.e. data are added linearly to the model). Only one model is fitted for a large range of weather situations thanks to the atmosphere dynamics being explicitly carried by NWPs. The result is a computationally inexpensive and easy way to include weather forecasts in PVPF models. Several references in the literature (e.g. [55, 61]) highlight that the use of NWPs as regressor features improves short-term forecasting performances in comparison with models fitted only with past production observations.

#### 1.5.3.2 Implicit integration

The alternative way to integrate weather information in a forecasting model is to consider it as a state variable. In this paradigm, the weather information acts as a kind of classification tool which associates PV generation data observed under similar atmospheric states. The underlying assumption behind this approach is that similar PV production dynamics are observed under similar weather dynamics. From a mathematical point of view, weather information is included in a nonlinear way to perform local regression (i.e. the model is trained on a coherent data subset with similar weather-characteristics to the expected weather situation). To this end, a similarity metric must be defined to measure the

likeness between two meteorological states. This approach provides a set of expert models dedicated to specific atmospheric states and is adaptive in the sense that the training of the model is conditioned to the weather situation. The terminology Weather-Conditioned (WHCO) is employed to refer to an approach, that operates a weather-based selection or classification in its learning dataset.

**Main Research Gaps - *Weather conditioning***

This weather conditioning strategy offers the possibility to condition several types of forecasting models, such as Auto-Regressive (AR) [62], ANN [63], or Support Vector Machines (SVM) [64, 65] models. To provide statistical information regarding the uncertainty of the forecast, conditioning methods can also be coupled with probabilistic approaches: [66] proposes a Quantile Random Forest (QRF) model trained on the 30 most similar days, while [67] derives probability distributions by applying a weighted kernel density estimation model on the most similar PV production observations. The literature highlights that the WHCO models exhibit higher forecasting skills than their counterparts trained on all past observations. Similar conclusions are drawn when NWPs are considered as explanatory features. However, to the best of the authors' knowledge, no studies have compared the two integration modes. This raises the following questions:

1. Between the explicit and implicit integration modes, which one expresses the full potential of weather information?
2. With the use of nonlinear models — which are able to consider a wide range of dynamics — is WHCO still relevant?

### 1.5.4 Spatio-temporal information

In the context of WHCO, spot NWPs data are usually used (i.e. weather parameters predicted at the nearest grid point of the plant's location). Such data allow us to work with few inputs but mainly reflect the temporal evolution of local weather conditions without providing information regarding their spatial characteristics (e.g. cloud distribution). As a result, it seems difficult for the WHCO strategy to efficiently take advantage of ST information. To fill this gap, [62] consider a WHCO approach based on forecasts of wind direction at the site location to select relevant ST information.

**Main Research Gaps - *Local or synoptic conditioning features***

The approach proposed in [62] is valid if the cloud motion remains linear, which can be assumed for very short lead times but can be contested for higher periods of time and high spatial scales. In the context of precipitation forecasting [68], large-scale circulation patterns represented by geopotential fields , namely gridded NWPs (i.e. two-dimensional data), are used to select situations with similar precipitation dynamics owing to their proven influence over cloud generation. Inasmuch as one can derive pressure gradients that drive air flow from high to low pressure regions from these fields, can these fields be used to provide a set of observations sharing temporal and spatial consistency and thus to improve the PVPF performances of models based on spatially distributed information?

## 1.6 Motivations and contributions of the thesis

The main motivations of this thesis have been touched upon in this preliminary chapter. In a few words, several flexibility solutions are investigated throughout the literature to deal with the issues raised by a shift in production means. These include storage systems [69] used to mitigate RES production variability and flexible loads [70] employed to adjust consumption to production (also called *demand response*). In comparison, RES forecasting appears as a cost-effective option. Wind generation forecasting is a widely studied area of research, to such an extent that it is considered as mature by the scientific community [49]. The situation is quite different regarding the PVPF realm, as illustrated by the gaps observed above.

To preserve grid stability, relevant forecasting tools are vital to counter the intermittency induced by the ever-increasing shares of renewables in the energy mix. This justifies the overarching goal of this thesis, which consists in improving the accuracy of short-term photovoltaic generation forecasts. To reach this objective, two main strategies have been identified: (1) to combine several sources of information; and (2) to extend existing statistical models. The first approach consists in analysing and exhibiting any spatio-temporal correlations that exist within the inputs, while the second option aims at studying the coupling of physics-based models with regression models as well as operating a shift from static to adaptive models. These two lines of research are associated with various scientific gaps that need to be overcome. A graphic summary can be seen in Figure 1.9. In addition, these scientific gaps ramify into secondary research gaps spread through the following chapters. To assist the readers, a summary is provided in Table 7.1 in the concluding chapter.

The main contributions related to the above-mentioned gaps are listed below:

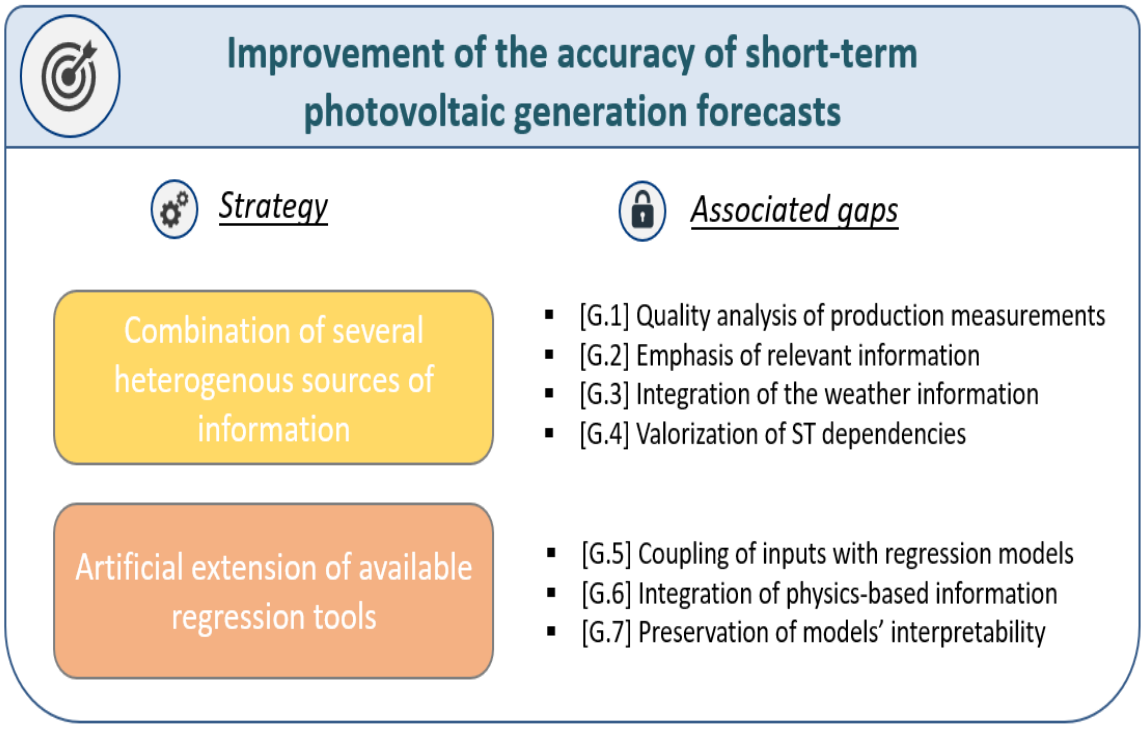



Figure 1.9 – Research gaps investigated throughout the thesis.

1. In the literature, there is a clear dichotomy between statistics-based and physics-based forecasts. Little permeability is observed on the grounds of the assumed ability of advanced regression algorithms to infer, on their own, the physical characteristics of the process. With an ambition to link these two fields, we evaluate the influence of considering power-like features instead of irradiance-related variables. Thus, this contribution covers research gap [G.6] and helps address [G.7].
2. An in-depth quality analysis of PV production is proposed in response to the research gap [G.1]. This approach aims at identifying and correcting production measurements associated with abnormal behaviours or component failures. This identification is based on a clustering algorithm coupled with a temporal segmentation approach. To the authors' knowledge, the proposed method is unique and no similar approaches have been applied within the PVPF domain, despite its impact on forecasting performances.
3. Special attention is paid to ST information. Clear-sky based normalisation is investigated to remove the deterministic component of the irradiance-based signal to facilitate correlation identification (i.e. research gap [G.2]). A novel method tackles the issue of the high dimensionality of Satellite Derived Surface Irradiance (SDSI), and has been the subject of a conference publication in [71]. The integration of satellite-based information is investigated and compared through several modelling strategies (this contributes to research gap [G.4]).
4. We propose a generic forecasting methodology to objectively condition any forecasting models to weather parameters. This work was published in a conference article [72]

and submitted to a journal. This directly addresses the research gaps [G.3] and [G.5]. This method is, for the first time, compared with the widespread approach consisting in considering weather information as additional explanatory variables.

## 1.7 Methodology

### 1.7.1 Architecture of the modelling chain

Rather than proposing an umpteenth brand-new model, the distinctive feature of this thesis is to focus on the different methodologies observed in the literature and to deduce best practices. This project has an iterative structure in so much as its constituent elements can be considered separately by interested readers. An overview of the modular architecture is given in Figure 1.10. In essence, this forecasting architecture is composed of a set of inputs (i.e. past production measurements, satellite-based information and NWPs) that are pre-processed before feeding a forecasting model to derive PV production forecasts at lead time $t + h$. Data sources are introduced in Section 2.4. The quality analysis of production observations is performed in Sections 4.2-4.3, the physical modelling of irradiance is the subject of Chapter 3, feature selection approaches are described in Sections 2.2, 5.2.4.2, and 5.3.2, and the normalisation process is described in Section 4.4. Last, statistical models are presented in Section 2.2, and the state conditioning model is described in detail in Chapter 6. To guide the reader, the narrative will rely on this graph, and zoom in at appropriate moments.

The main advantage of such an architecture is its flexibility: depending on the characteristics investigated, it is possible to activate or put to sleep specific blocks. In addition, the set of features can be easily extended according to needs. Within the scope of this study we will focus on two state-of-the-art regression models; nevertheless, block number 5 makes it possible to use this modelling strategy with a wider range of models.

### 1.7.2 Evaluation framework

To achieve higher forecasting skills, new forecasting architectures and information preprocessing steps have been investigated throughout this thesis. To be retained in the whole modelling chain, each new process developed needs to justify an enhancement. At this point it is legitimate to ask how respective improvements can be evaluated.

#### 1.7.2.1 Preprocessing steps validation scheme

Two alternatives have been considered regarding the validation of data preprocessing steps (i.e. steps 1, 2, 3, and 4 in Figure 1.10).

The first option consists in assessing the predictability of the preprocessed time series. To do so, entropy-based indicators, such as Approximate Entropy (ApEn) and Sample Entropy (SampEn) from the information theory domain, are usually used [73]. These parameters

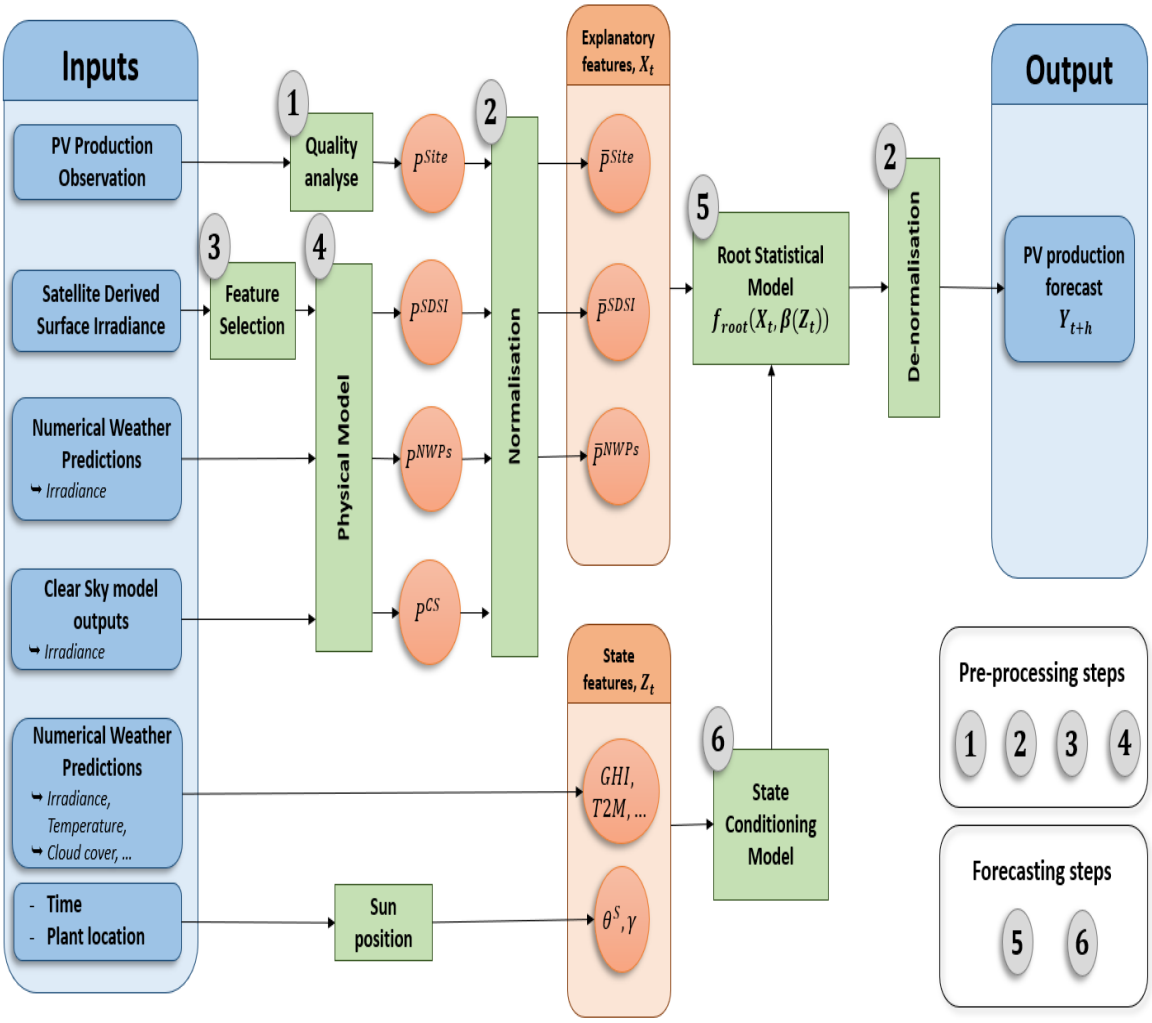



Figure 1.10 – Main architecture of the forecasting chain.

characterise the predictability of a time series by describing its complexity and its degree of self-similarity in terms of patterns. This *a priori* approach is appealing in a context of a high number of time series to test.

The second option, which is also more computationally expensive, determines the forecasting enhancement induced by data preprocessing steps through an *a posteriori* approach. The latter considers the use of a regression model to forecast PV production from preprocessed features. Within this paradigm, relevant preprocessing steps are associated with the most accurate forecasts.

An empirical comparison between both approaches highlighted that entropy-based indicators may provide misleading conclusions (i.e. contradiction between entropy-based results and forecast performances). This is assumed to result from a poor choice of parameters. This motivates the use of the second option.

#### 1.7.2.2 Forecasting performance assessment

We observe in the literature a growing demand to provide reproducible results associated with guidelines and frameworks for evaluating the quality of forecasts [45, 74, 75].

As forecasts are data-, location-, and time-step-, dependent, it is necessary to provide data and codes to allow reproducible analysis. Unfortunately, as this project involves an industrial partner, data and code sharing is not permitted. However, a special effort will be made by the author to detail as much as possible the approaches developed and relevant references will be provided throughout this study.

To comply with forecasting good practices, first, it is necessary to perform comparisons with comparable elements. In [75], the author points out that conventional metrics alone cannot be used to compare forecasts carried out with different datasets, locations or horizons. Instead, such scores should be used in the perspective of a forecasting benchmark via a skill score. A consistent benchmark approach found in several studies is the smart persistence (i.e. a persistence model using the forecast Clear-Sky Index (CSI)) [9]. This skill score is computed with a set of common evaluation metrics, namely, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Mean Bias Error (MBE).

### 1.7.3 A real case study

The modelling architecture is built taking operational constraints into consideration, while the evaluation framework is performed with real datasets.

Apart from PV production measurements of the CNR fleet, all data involved in this thesis are issued by research institutes (NWPs are generated by the European Centre for Medium-Range Weather Forecasts (ECMWF)) or private companies (TRANSVALOR S.A provides SDSI) that are known actors in the PVPF field.

The model developed in this PhD aims at an operational application. As such, it seems relevant to provide the computational time (Appendix A) needed to perform model fitting and testing. Since such data are highly hardware-dependent, it is necessary to consider our model in relation to a specific model.

## 1.8 Structure of the thesis

This opening chapter presents the main challenges associated with a high penetration of renewable energy in the electrical grid and introduces the subject of PV generation forecasting. Far from having disclosed all its secrets, this subject is still very active and several research gaps have been identified. In the following chapters, we attempt to provide some answers. To guide the reader, a roadmap depicting the structure of the different chapters is presented in Figure 1.11. This graph details the methods and inputs used to

9. The output obtained by normalising irradiance-related features by the irradiance observed in a cloudless sky. Such quantity possesses stationary-related properties.

elaborate the preprocessing and forecasting steps introduced in Figure 1.10 as well as the associated main research gaps presented previously.

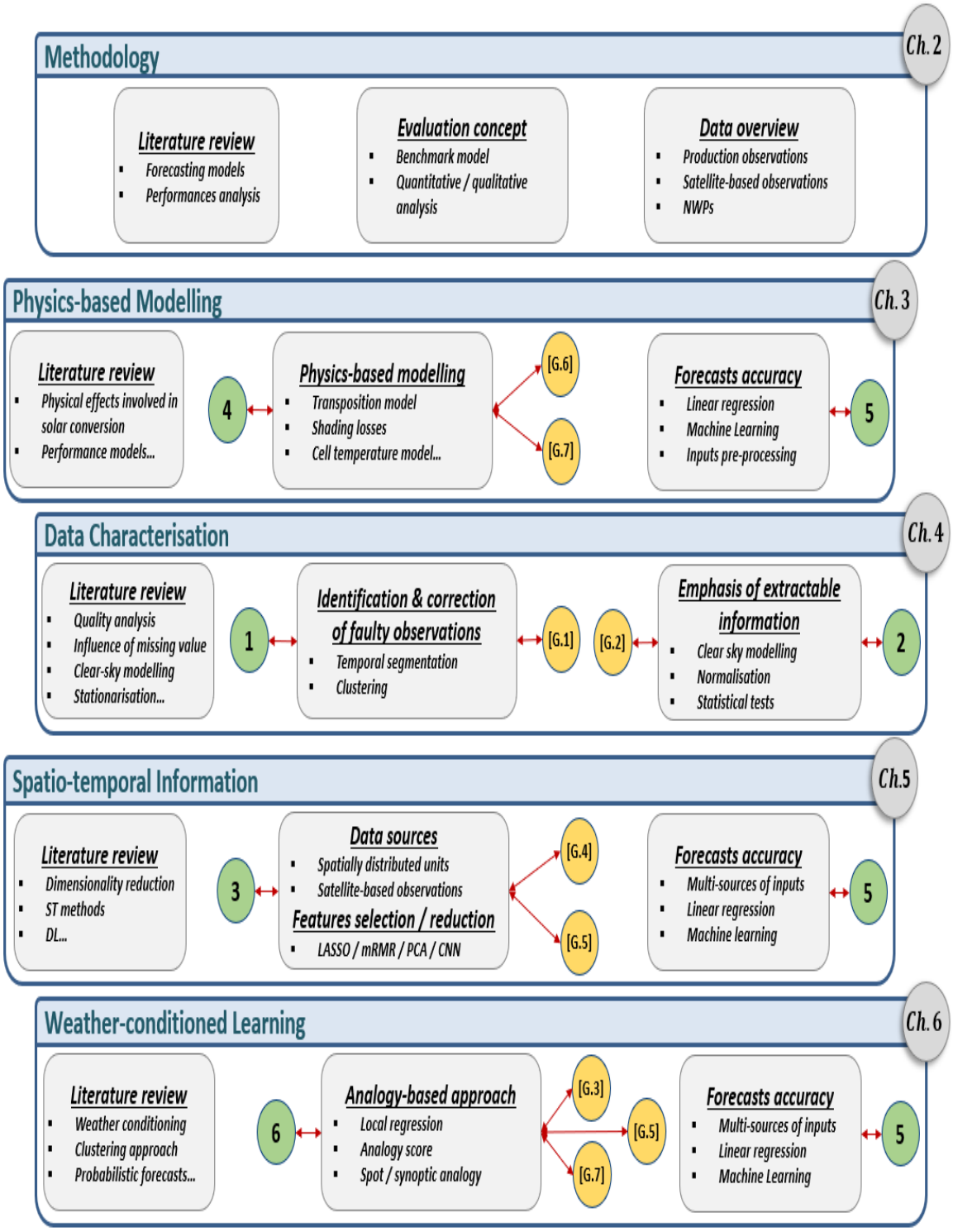



Figure 1.11 – Graphical outline of the structure of the thesis. Grey boxes detail the methodology (e.g. methods, algorithms, inputs) adopted within the chapters to answer the main research gaps (Figure 1.9) represented by the yellow items, while green items represent modular blocks numbering in Figure 1.10.

The remainder of the document is organised as follows:

Throughout this manuscript, various investigations are carried out to design forecasting models with greater capacities than those found in the literature. This will raise two main questions (1) what do we want? and (2) how do we measure it? Several criteria exist to compare forecasting models. In Chapter 2, the focus is on performance and more specifically on accuracy. Recommended standard practices observed in the literature to evaluate forecasts are implemented to enable a precise quantification of accuracy and to allow comparison with other studies. After that, the different sources of inputs are introduced and characterised.

Chapter 3 delves into the physical modelling of the conversion processes involved in producing electricity from solar resources. The different steps are described, ranging from the irradiance crossing the atmosphere to the increase in the voltage level performed by the transformer. The technical characteristics of PV panels are also considered (e.g. modules orientation, temperature response of the cells). This aims at deriving a physics-based model that converts solar irradiance into electrical power from physics-based knowledge. This performance module is then integrated into the forecasting architecture developed so far as a preprocessing step before the regression model. The main ambition behind this methodology is to reduce the computational effort of the regression model by integrating known information. This methodology is compared with the direct integration of raw information (i.e. without the use of the physics-based model) in order to judge the statistical model's capacity to derive knowledge from its learning.

Chapter 4 proposes an in-depth analysis of the features used to forecast PV generation. First, particular attention is paid to the reliability of PV production observations. Since PV parks are composed of a set of conversion devices, the failure of one of the latter deteriorates the quality of the production signal: in this case, production level variations result from technical issues rather than meteorological variations. An identification and imputation strategy is developed to filter and potentially correct abnormal behaviour (e.g. curtailment, components shutdown). Second, irradiance-based features (i.e. PV production, SDSI, or Surface Solar Radiation Downwards (SSRD)) are cleaned from their deterministic component. The latter is linked with the Sun's movement within the sky dome. Such an approach shines a light on the part of the signal associated with cloud movement, and consequently improves ST dependencies between features. Last but not least, this process improves the stationary properties of the time series, a requirement for the use of specific statistical tools.

The use of ST information, namely spatially distributed PV units and satellite-based information, is developed in Chapter 5. First, the potential of Spatially Distributed Units (SDU) information from the CNR network is investigated. A physics-based selection of sites and relevant lags is proposed in line with prevailing winds. However, the north-south distribution of sites contrasts with the east-west orientation of dominant winds. This legit-

imates a resort to satellite-based information to freely select spatial observations. In this regard, a novel method is introduced to perform relevant feature selection to maximise information relevance while reducing the computational burden induced by satellite-derived observations. Given that such a variable is well suited for Convolutional Neural Networks (CNN) architecture (i.e. due to its two-dimensions), we investigate a preprocessing method aiming at establishing irradiance forecasts at the site location from the sequence of the last maps. Finally, satellite maps derived from infra-red channels are included in the forecasting architecture with the target of improving forecasts for the early morning.

Chapter 6 proposes a generic methodology to condition the learning of regression models to the weather state and to obtain weather-specific models. Such an approach can be viewed as a way to make models adaptive by dynamically updating their parameters. These expert models, based on local regression, are derived from the analogy principle widely used in the meteorological domain. The conditioned learning is performed alternatively with spot data (i.e. predictions at the park position) and gridded data (i.e. geopotential fields). This aims at investigating how the nature of weather information influences ST dependencies. Lastly, a forecasting performance comparison between conditioned and un-conditioned models fed with several kind of explanatory features provides guidelines regarding the appropriate family of models to use. The properties of probabilistic forecasting derived with WHCO approach are also assessed.

Finally, Chapter 7 draws the main conclusions of this thesis as well as potential future research opportunities.

## 1.9 List of publications, conferences and presentations

The present thesis led to the following publications:

### 1.9.1 Peer-reviewed journal

1. Bellinguer K., Girard R., Bontron G., Kariniotakis G., A Generic Methodology to Efficiently Integrate Weather Information in Short-term Photovoltaic Generation Forecasting Models (Accepted for publication in Solar Energy).
2. Bellinguer K., Girard R., Bontron G., Kariniotakis G., Short-term Photovoltaic Power Forecasting Enhanced by Heterogeneous Sources of Spatio-temporal Data (Submitted for publication).

### 1.9.2 Conference papers

1. Bellinguer K., Girard R., Bontron G., and Kariniotakis G., Short-Term Photovoltaic Generation Forecasting Using Multiple Heterogenous Sources of Data. In *36th European Photovoltaic Solar Energy Conference and Exhibition*, Sep 2018, Marseille, France, https://hal.archives-ouvertes.fr/hal-02314083.

2. Bellinguer K., Girard R., Bontron G., and Kariniotakis G., Short-term Forecasting of Photovoltaic Generation based on Conditioned Learning of Geopotential Fields, 2020 *55th International Universities Power Engineering Conference (UPEC), Turin*, Italy, 2020, pp. 1-6, https://hal.archives-ouvertes.fr/hal-02932018.

3. Bellinguer K., Girard R., Bontron G., and Kariniotakis G., Short-Term Photovoltaic Generation Forecasting Enhanced by Satellite Derived Irradiance. *26th International Conference & Exhibition on Electricity Distribution (CIRED 2021)*, CIRED, Sep 2021, Virtual Event, Switzerland, https://hal.archives-ouvertes.fr/hal-03407898.

### 1.9.3 Conference presentations

1. Bellinguer K., Girard R., Bontron G., and Kariniotakis G., Short-term photovoltaic generation forecasting using multiple heterogenous sources of data based on an analog approach., *EGU General Assembly 2020*, Online, 4–8 May 2020, EGU2020-13790, https://doi.org/10.5194/egusphere-egu2020-13790.

2. Bellinguer K., Girard R., Bontron G., and Kariniotakis G., Assessment of Alternative Ways to Integrate Weather Predictions in Photovoltaic Generation Forecasting., *EGU General Assembly 2021*, online, 19–30 Apr 2021, EGU21-16091, https://doi.org/10.5194/egusphere-egu21-16091.

### 1.9.4 Additional communications

The 17$^{\text{th}}$ International Conference on the European Energy Market EEM20 set up a competition to develop probabilistic forecasting tools of wind production at a regional level. Our team, composed of Valentin MAHLER, Simon CAMAL and myself from Mines Paristech, proposed a model that won the competition and led to a conference paper, and two presentations at ISF20 and IEA Wind Forecasting [10]:

- Bellinguer, K., Mahler, V., Camal, S., and Kariniotakis, G., Probabilistic Forecasting of Regional Wind Power Generation for the EEM20 Competition: a Physics-oriented Machine Learning Approach, 2020 *17th International Conference on the European Energy Market (EEM)*, Stockholm, Sweden, 2020, pp. 1-6, https://hal.archives-ouvertes.fr/hal-02952589.

- Bellinguer K., Mahler V., Camal S., and Kariniotakis G., Forecasting regional wind production based on weather similarity and site clustering for the EEM20 Competition. *40th International Symposium on Forecasting – ISF20*. Virtual conference, October 2020, https://hal.archives-ouvertes.fr/hal-03157849.

10. https://www.youtube.com/watch?v=n0m3Sl8Zwtk

## 1.10 Résumé en Français

### Contexte

Depuis la révolution industrielle, les émissions de gaz à effet de serre n'ont cessé d'augmenter du fait des activités anthropogéniques, conduisant ainsi au réchauffement climatique. Pour faire face à cette menace pressante, de profonds changements de nos modes de production énergétique sont nécessaires. L'une des principales sources de gaz à effet de serre étant la production d'électricité, les énergies renouvelables constituent une alternative louable au gaz et au fioul. En 2015, se tenait la 21ème conférence des parties sur les changements climatiques. Malgré la laudation des accords de Paris, aucun mécanisme visant à assurer le respect des engagements établis n'a été défini si bien que chaque pays est libre de définir ses propres objectifs. En ce qui concerne la France, plusieurs feuilles de routes (e.g. la Stratégie Nationale Bas-Carbone (SNBC)) ont permis de définir les orientations à mettre en œuvre afin d'assurer le respect des objectifs introduits dans les textes réglementaires tels que la Loi de Transition Energétique pour la Croissance Verte (LTECV).

De par son histoire, la France repose aujourd'hui pour une part importante sur l'énergie nucléaire. Ce parc se faisant vieillissant, la France a exprimé, au travers de la LTECV, sa volonté de réduire la part du nucléaire dans le mix énergétique à 50% d'ici 2035, au profit de l'augmentation des renouvelables. Ainsi, pour promouvoir leurs installations, des dispositifs d'aide ont été mis en place par les différents gouvernements si bien que l'on observe une croissance soutenue du PV (Figure 1.2) depuis 2010.

Dans nos pays occidentaux, l'électricité est considérée comme acquise pour bon nombre d'entre nous. Pourtant, à chaque instant des efforts colossaux sont déployés pour assurer son approvisionnement et l'utilisation en toute sécurité des moyens de production et de consommation. A l'heure actuelle, il est difficile de stocker l'électricité à grande échelle, de ce fait il est nécessaire d'assurer un équilibre parfait entre la consommation et la production. On comprend alors que l'intégration à grande échelle des moyens de production renouvelable, qui sont par nature intermittent, occasionne d'importantes contraintes quant à la stabilité du réseau.

Outre l'augmentation de la part des renouvelables, la transition énergétique s'accompagne d'une véritable transformation des usages, des services et des acteurs. On observe l'électrification des secteurs fortement émetteurs en gaz à effet de serre : on peut penser par exemple au secteur du transport avec le développement des véhicules électriques. Ces mêmes véhicules seront en mesure de fournir des services au réseau électrique en y injectant de l'énergie pour faire face au pic de consommation. Dans ce contexte, le consommateur peut devenir un prosommateur en injectant l'énergie produite par exemple par les panneaux PV de son habitation, ou il peut également rendre des services au réseau en autorisant l'arrêt automatique de certains appareils lors des pics de consommation.

Pour assurer le bon fonctionnement du réseau, les opérateurs doivent maintenir l'équi-

libre entre la production et la consommation. De ce fait, des pénalités financières sont imputées aux responsables d'équilibres qui dérogent à cette règle au niveau de leur périmètre d'équilibre. De plus, les mécanismes de support dédiés aux producteurs arrivent à termes dans de nombreux pays européens, ce qui contraint ces derniers à intégrer les marchés financiers de l'énergie. Dès lors les producteurs soumettent des offres concernant les quantités d'énergie qui seront disponibles à la vente. Ils se doivent alors de respecter leurs engagements au risque de payer des pénalités financières.

Deux types de technologies sont typiquement utilisés pour produire de l'électricité à partir de la lumière du soleil : (1) les systèmes PV, et (2) les systèmes solaires thermodynamiques à concentration. Cette dernière technologie utilise des lentilles ou des miroirs pour concentrer la lumière du soleil et chauffer un fluide qui va activer une turbine à gaz ou directement alimenter un processus industriel. Les systèmes PV quant à eux reposent sur le principe photovoltaïque qui permet de convertir l'irradiance solaire en courant électrique continu. Aujourd'hui, les systèmes PV représentent la part de marché la plus importante. Même si le taux de pénétration du PV reste modeste en France (i.e. de l'ordre de 2.5%), ce dernier est supposé augmenter jusqu'à 9% - 37% en 2050 selon les différents scénarii de RTE.

La production PV est caractérisée par une forte variabilité mais également par une prédictibilité limitée en raison de la nature chaotique de l'atmosphère. La variabilité de la production PV peut être décomposée en une composante déterministe, résultat du mouvement du soleil qui induit une variabilité journalière et saisonnière, et une composante stochastique qui est le fruit des mouvements de masses atmosphériques.

### Etat de l'art

Dans la littérature plusieurs options sont étudiées afin de limiter les effets négatifs de la variabilité de la production PV sur le réseau. A titre indicatif, on peut citer le développement de systèmes de stockage d'énergie tels que les batteries électriques. Dans le cadre de ce doctorat, nous avons fait le choix de nous focaliser sur le domaine de la prévision court-terme de la production (i.e. de 15 minutes à plus 6 heures). L'objectif premier de ce sujet de recherche est de proposer une approche de prévision (1) qui soit aussi simple que possible afin de permettre son utilisation par le plus grand nombre, (2) qui puisse être utilisée en temps réel (ce qui suppose une méthode robuste et rapide à caler), et enfin (3) qui puisse être extensible dans le sens où elle permet l'addition de nouvelles variables. En dépit du fait qu'à l'heure actuelle les grands axes de recherche sont orientés vers des modèles probabilistes, nous avons fait le choix de nous intéresser aux modèles déterministes et ce pour plusieurs raisons. Par exemple, les modèles déterministes permettent de comprendre aisément les relations entre les données d'entrée et la sortie du modèle. Par ailleurs, l'analyse de la précision des prévisions est plus simple que dans le cas probabiliste.

Traditionnellement, les modèles utilisés dans le domaine de la prévision PV se décom-

posent en trois groupes : les modèles physiques, les modèles statistiques et une combinaison des deux précédentes approches. Les modèles physiques peuvent être vu comme une boîte blanche dans laquelle le processus de conversion de l'irradiance en puissance électrique est explicitement modélisé. D'un autre côté, les modèles statistiques eux ne présupposent aucune connaissance a priori du processus, bien au contraire, ils infèrent les lois de conversion à partir des données historiques. Cette dernière approche est majoritairement représentée dans la littérature mais pose le problème de l'interprétabilité des résultats : les modèles, de par leur complexité, sont assimilés à des boîtes noires et la relation établie entre l'entrée et la sortie peut être plus ou moins opaque. A ce stade, on peut se demander si l'inclusion de connaissance physique dans les modèles statistiques peut avoir un effet positif sur la qualité des prévisions.

D'un autre côté, depuis quelques années déjà, nous observons une tendance qui consiste à combiner des sources de données hétérogènes. Sur la plage d'horizon qui est la nôtre, il est commun de considérer les dernières mesures de la production en tant que variables explicatives de la production future. Un parc PV est un système complexe composé de plusieurs sous-composants. Ces différents éléments sont à même de subir des avaries et donc de réduire artificiellement la production de manière plus ou moins aléatoire. Il est alors pertinent de s'interroger quant à l'impact de ce signal de production détérioré sur les performances prédictives des modèles. Bien loin de se cantonner à cette source de données, la littérature considère également des observations de la situation atmosphérique pouvant revêtir diverses formes. Par exemple, pour des horizons très court-termes, des observations sur site obtenues à partir de caméras hémisphériques sont généralement utilisées. Les observations de la couverture nuageuse peuvent également être obtenues depuis le ciel via des satellites. Cette source d'information possède l'avantage de couvrir une zone spatiale conséquente qui peut atteindre plusieurs centaines de kilomètre mais ce, au détriment de la résolution spatiale qui est usuellement de l'ordre de quelques kilomètres. Etant donnée la forte dimension des images satellite, des méthodes de sélection ou de réduction doivent être mise en place. Ce type d'information est généralement utilisé pour des horizons allant de quelques minutes à environ 6 heures. Pour des horizons plus importants, nous nous tournons vers des prévisions numériques du temps obtenues à partir de modèles physiques simulant l'atmosphère.

Ces prévisions numériques peuvent être intégrées dans les modèles de prévisions selon deux approches distinctes. La première considère les données NWPs comme des variables exogènes. Dans ce cas, la dynamique atmosphérique est explicitement portée par les variables NWPs. La seconde approche considère quant à elle les données NWPs en tant que variable d'état. Dans ce paradigme, les données NWPs agissent comme une variable de classification et permettent l'obtention de modèles experts dédiés à certains types de dynamiques atmosphériques. Selon cette approche l'information pertinente est portée par les variables explicatives. La littérature montre que ces deux approches offrent de meilleures performances en comparaison à un modèle uniquement basé sur l'historique de production.

Toutefois, à notre connaissance, aucune étude ne les compare simultanément. D'autre part, cette stratégie de conditionnement par la situation météorologique est appliquée à une large gamme de modèles de régression, néanmoins, on peut s'interroger quant à la pertinence d'appliquer ce type d'approche à des modèles complexes tels que les modèles non-linéaires.

Dans un contexte de conditionnement par la situation météorologique, des prévisions localisées au niveau du site d'intérêt sont utilisées. De telles données offrent l'avantage de travailler avec un nombre limité de variables mais ne permettent pas de refléter les caractéristiques spatiales de la situation atmosphérique aux alentours du site (e.g. couverture nuageuse). Ainsi, le conditionnement tel qu'on le trouve dans la littérature ne semble pas être à même de valoriser l'information ST contenue dans des données telles que les observations obtenues par imagerie satellite. Dès lors, on peut se demander si des variables d'état telles que les champs géopotentiels peuvent nous aider à caler des modèles ST en nous fournissant des situations qui partagent des similarités tant en termes d'évolution temporelle que spatiale.

En résumé, nous avons identifiés les questions de recherche suivantes :

- **RQ1 :** Comment les défauts des composants principaux des centrales PV ont des répercussions sur la précision des prévisions ?
- **RQ2 :** Quelle est la meilleure approche pour mettre en valeur l'information pertinente contenue dans les séries temporelles de production ou d'irradiance ?
- **RQ3 :** Quelle est la stratégie optimale pour coupler plusieurs sources d'information ?
- **RQ4 :** Quelle stratégie pour gérer des jeux de données de grande dimension ?
- **RQ5 :** Quelle est la meilleure approche pour intégrer des connaissances physiques dans un modèle statistique ?
- **RQ6 :** Comment pouvons-nous améliorer l'interprétabilité des modèles boîtes noires ?

### Contributions de la thèse

L'objectif principale de ce sujet de recherche consiste en l'amélioration de la précision des prévisions de la production PV. Pour ce faire, notre stratégie repose sur deux points essentiels : (1) combiner plusieurs sources hétérogènes d'information, et (2) étendre de manière artificielle les modèles de régression. Au regard de la littérature nous proposons les contributions suivantes :

1. Nous observons une dichotomie assez marquée entre les modèles dérivés des statistiques et les modèles obtenus à partir des connaissances physiques. Cette faible perméabilité peut être expliquée par la croyance en les capacités du modèle statistique à inférer de lui-même les caractéristiques du processus physique. Avec l'ambition de relier ces deux domaines, nous proposons de prétraiter l'irradiance afin de la convertir en une puissance électrique via un modèle de conversion physique qui intègre un nombre restreint de paramètres physiques.

2. Une analyse approfondie de la qualité des données de production est proposée. Cette approche vise à identifier et à corriger des observations associées à une anomalie de production. L'identification est basée sur un algorithme de clustering couplé à une segmentation temporelle alors que le processus de correction se base sur nos connaissances de l'architecture de la centrale PV. A notre connaissance, cette approche est unique dans le domaine de la prévision PV.
3. Une attention toute particulière est portée aux données ST. La normalisation par ciel-clair est étudiée afin de supprimer ou tout du moins réduire l'influence de la composante déterministe de l'irradiance et faciliter l'identification de corrélations ST. Une nouvelle approche de sélection de variables est également proposée dans l'optique de réduire la dimension des données satellitaires. Contrairement à d'autres approches clés de la littérature, la méthode proposée permet la sélection d'un ensemble de pixels spatialement distribués, ce qui permet de minimiser la redondance de l'information sélectionnée. Enfin, des observations satellitaires obtenues à partir de canaux infrarouges sont intégrées pour leurs influences positives sur les prévisions générées pendant la nuit.
4. Nous développons les fondements mathématiques d'une méthodologie générique permettant de conditionner n'importe quels modèles de prévisions à la situation météorologique. Initialement utilisée avec des modèles déterministes, cette approche est également appliquée de manière préliminaire avec des modèles probabilistes.

### Structure de la thèse

La suite de ce manuscrit est articulée de la manière suivante :

Le Chapitre 2 introduit la méthodologie adoptée tout au long de cette étude, à savoir les modèles considérés, les diverses données d'entrée et également les critères d'analyse et de comparaison des prévisions. Une étude préliminaire de notre jeu de données est proposée et comparée à la littérature.

Le Chapitre 3 se plonge dans la modélisation physique du processus de conversion de l'irradiance en puissance électrique. Les différents processus sont décrits alternativement et leurs influences sur les performances prédictives sont étudiées.

Ensuite, une analyse approfondie de la qualité des données de production est proposée au Chapitre 4. Tout d'abord, les données associées à une détérioration des systèmes de conversion sont identifiées et éventuellement corrigées selon les données à disposition. Non seulement ce chapitre cherche à effacer la variabilité artificielle induite par les avaries techniques, mais en sus il cherche à supprimer la variabilité déterministe due à la course du soleil.

L'utilisation de données ST est spécifiquement étudiée au Chapitre 5. Dans un premier temps, nous nous intéressons aux données fournies par le réseau de centrales à notre disposition. Etant donnée l'inadéquation entre la distribution spatiale de ces sites et la distribution

des vents dominants, nous nous tournons vers des données d'origine satellitaire. Deux types de données sont considérés : (1) des estimations de l'irradiance au sol, et (2) des observations de l'opacité nuageuse obtenues par canaux infrarouges.

Le Chapitre 6 quant à lui développe une méthodologie permettant de conditionner un modèle de régression à la situation météorologique. Cette approche permet de rendre le modèle adaptatif. Le conditionnement est réalisé soit avec des données localisées au niveau du site d'intérêt, soit via des données 2D, en l'occurrence un champ géopotentiel. Ce chapitre propose des recommandations concernant l'utilisation de telle ou telle variable selon le modèle considéré. Pour finir, la méthodologie de conditionnement est appliquée sur des modèles probabilistes.

Enfin, le Chapitre 7 tire les principales conclusions de cette thèse et propose de nouveaux axes de recherche.

# Chapter 2

# Forecasting Methodology

All models are false, but some are useful.

George Box (1979)

**Contents**

## 2.1 Introduction

At least two directions are conceivable to improve forecast accuracy: (1) extend the inputs to integrate more explanatory information on the process, or (2) reduce the model uncertainty.

The former aspect is specifically tackled in Section 4.2 jointly employing several sources of information. Due to physical constraints and modelling limitations, today it is not possible to directly use Numerical Weather Predictions (NWPs) model outputs with enough accuracy to forecast short-term irradiance at the precise location of a power plant. To fill this gap, the great majority of forecasters turn to data-driven approaches, and increasingly rely on several sources of information to extend the range of available information.

Selecting the optimal model for a specific application is not an easy task, as none outperforms the others in all conditions, and a profusion of models is proposed in the Photovoltaic Production Forecasting (PVPF) domain, each with its own strengths and weaknesses [46, 76, 77]. To deal with this issue, we consider a representative approach from the two mainstream classes of parametric regression models, namely, the linear and nonlinear family of models. To guide our selection, we consider criteria related to complexity, interpretability, and scalability [1]. An extensible modular framework depicted in Figure 1.10 is built throughout this document. Initially based on state-of-the-art regression models, we extend this framework with new modules in an incremental way, which allows us to increase the complexity of the forecasting strategy. Lastly, we adopt a validation framework recommended by the literature.

This chapter presents the forecasting models used to generate production forecasts as well as the validation framework implemented to assess their quality. Then, the different sources of information considered as inputs are introduced, and an implementation of the forecasting and validation frameworks is provided. The general workflow of this chapter is displayed in Figure 2.1.

## 2.2 Forecast generation

### 2.2.1 Baseline forecasting framework

The main objective of this work is to predict future Photovoltaic (PV) power generation, $h$ minutes ahead, based on fusing temporal and spatial information at the targeted site and its neighbourhood.

Figure 2.2 depicts the modular architecture used as a baseline to forecast PV production. This architecture is progressively extended throughout the course of this narrative. The fundamental modules implemented in this forecasting architecture are first the root model, and second the (de)-normalisation model. The root model represents the data-driven framework

1. Scalability refers to the property of a model to handle an increasing amount of data.

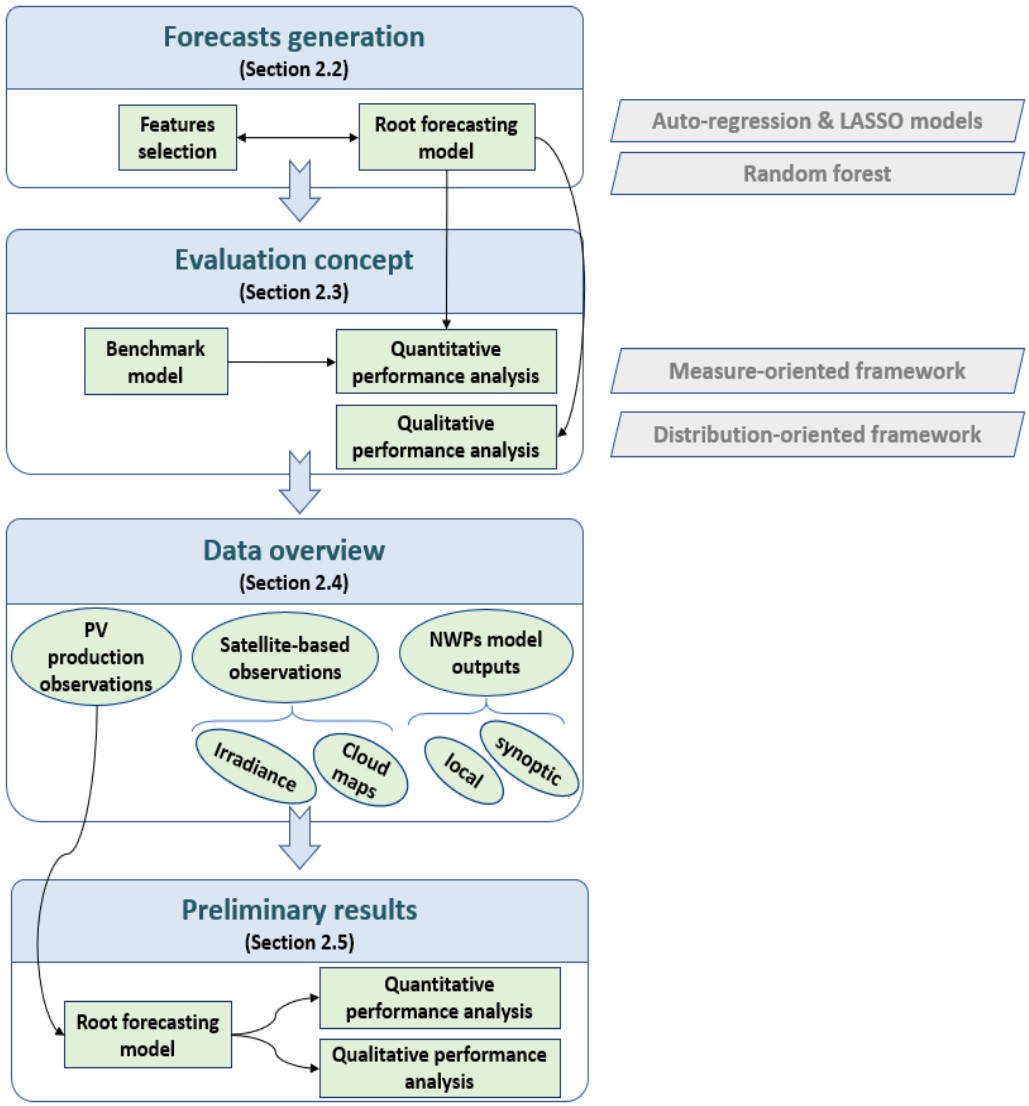



Figure 2.1 – General workflow of the chapter.

used to derive the forecast at time $t+h$ from a set of regressors. A generic formulation is provided by the following equation:

$$Y_{t+h|t} = f_{root}\left(X_t, B^h\right) + \epsilon_t. \tag{2.1}$$

| | |
|---|---|
| $Y_{t+h\|t}$ | Vector of the response variable at time $t+h$, |
| $f_{root}$ | Root regression model employed for the mapping of $X_t$ to $Y_{t+h}$, |
| $X_t$ | Vector of explanatory features which may contain past production and satellite-derived observations as well as NWPs model outputs, |
| $B^h$ | Vector of the model's parameters to be estimated, |
| $\epsilon_t$ | Error term representing random errors or variability from sources not considered. |

The normalisation framework is used to remove trends observed within solar-related time series (e.g. PV production, Global Horizontal Irradiance (GHI) measurements) be-

fore the processing of information by the regression tools. In this work, we consider the clear-sky outputs normalisation approach. The latter consists of normalising a feature by the theoretical irradiance (or alternatively PV production) that would have been observed in a cloudless sky. This allows us to remove seasonal and diurnal trends of solar-related features resulting from the Sun's path. The output of this normalisation process is called the Clear-Sky Index (CSI). In short, this CSI possesses better stationarity properties than PV production, and highlights Spatio-temporal (ST) dependencies within features. The normalisation process is detailed later on in Chapter 4. Lastly, the de-normalisation process converts CSI predicted values back into predicted PV production values.

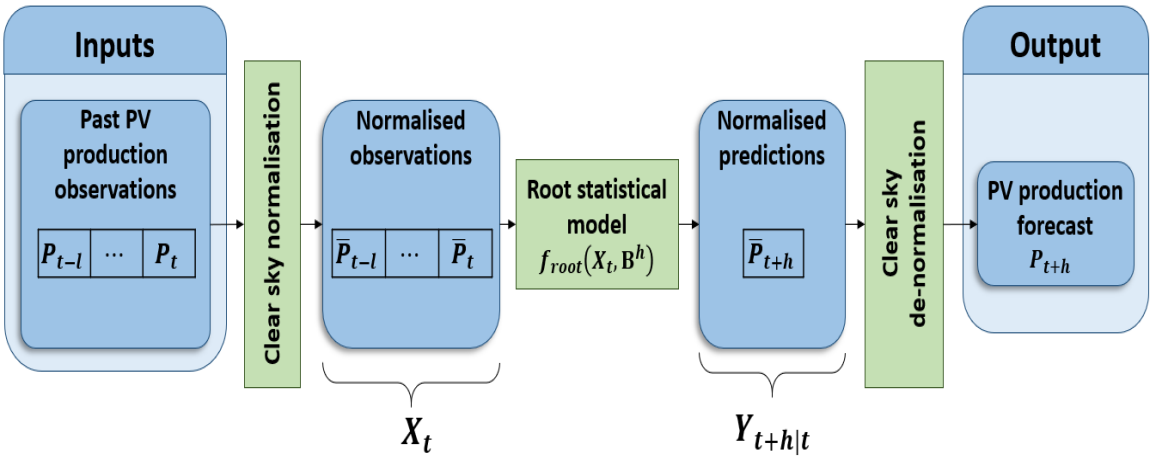



Figure 2.2 – Baseline forecasting framework. The normalisation process is detailed in Chapter 4.

All methods are implemented using R statistical programming language [78]. Despite R not being the most popular programming language (it was ranked 6$^{\text{th}}$ by IEEE Spectrum in 2017), it enables an easy interfacing with other programming languages such as Python-based deep neural network libraries.

### 2.2.2 Choice of the models

When dealing with forecasting, the question arises of which $f_{root}$ model to use. This choice may be motivated by model forecasting accuracy, computational costs, tuning complexity, maturity, and even interpretability. Often the best choice is a trade-off between these different options.

At first glance, interpretability appears to be an obscure concept that people refer to when they want to get an insight into the logic built by the model during its learning phase. Understanding how a model works is of prime importance, first, in order to check that it is working correctly, but also to identify areas for improvement. A classic approach to gain knowledge of the driving forces at work is to compute features importance. Features importance is a way to assess the predictive impact of each input and can be used to select the most relevant features or perform feature engineering. Intuitively, one can say that the more complex an algorithm is, the less interpretable it is. In this prospective project, we

made the choice to work with interpretable models as much as possible with an ambition to improve forecasting performances by combining heterogeneous sources of information, rather than increasing the root model's complexity. However, the approach we developed is general enough to provide conclusions that may be applied with more complex regression strategies.

### 2.2.3 Linear model

In the solar irradiance and PV power forecasting field, the Auto Regressive Integrated Moving Average (ARIMA) family [79] is the most widely used time series method [80]. To explain this, the authors point out that it is a common choice for a reference method. Yet, over the last ten years, this set of models has been used in a satisfying manner in the short-term PV generation forecasting field: [61, 62, 81] consider Auto-Regressive with eXternal inputs (ARX) models, while [82] proposes the Coupled Auto Regressive Dynamical System (CARDS) model, just to name a few. ARX [56, 58, 82, 83] and Vector Auto-Regressive (VAR) [2] [84, 85] models are also very present in the ST-related literature. The reasons for this success may be attributed to the relative good accuracy, rapid training, low complexity and maturity of this family of models. In this work, we chose to work with the Auto-Regressive (AR) model. An AR model was chosen rather than a VAR model because the former provides more flexibility, especially in terms of weather conditioning (notion defined in Chapter 6). In simple terms, a conditioned model aims at dividing the dataset into subsets of production levels observed under similar weather patterns. We understand that such an approach may only be used with VAR models if the whole power plant network is subject to the same weather conditions.

#### 2.2.3.1 Autoregressive model

In the present study, the AR model (Equation 2.2) is considered as the linear root model of our modelling strategy. This model provides easy-to-understand regression coefficients, which allow an in-depth forecasting performance analysis. Given its low complexity, the AR is not in a position to capture the broad range of PV generation dynamics associated with each weather state. Therefore, to extend its forecasting capacities, Amaro et al. turn to wind-conditioning in [62], whereas [61] proposes an adaptive model that favours most recent observations by applying a Recursive Least Squares (RLS) method.

$$f_{root}\left(X_t, B^h\right) = \beta_0^h + \beta^h X_t^\intercal \tag{2.2}$$

#### 2.2.3.2 Feature selection

An increasing number of explanatory variables makes the model more complex and can undermine its accuracy. To tackle this issue, the Least Absolute Shrinkage and Selection

2. Model specifically designed for the analysis of spatially sparse ST data [84].

Operator (LASSO) [86] (Equation 2.3) procedure is implemented in the AR model to perform feature selection and regularisation, while enhancing the interpretability of the model. During the determination of the regression coefficients, we add a term composed of a tuning parameter, called $\lambda$, multiplied by the sum of absolute values of the coefficients. This has the effect of forcing coefficients with minor contributions to be equal to zero. When $\lambda$ is zero, it simply gives the least squares fit, but as the parameter grows, more variables are set to zero. A k-fold cross-validation [3] approach performed in the training set is used to achieve the best tuning of the $\lambda$ parameter. In addition, this feature selection method proposes a non-parametric approach regarding the selection of the AR model order $L$ (i.e. the number of PV production lags to consider in the model). By default, we consider the production lags up to 2 hours, then, for each horizon, the optimal sets of regressor features is defined by the LASSO.

$$(\hat{\beta_0^h}, \hat{\beta^h}) = \underset{\beta_0^h, \beta^h}{\arg\min} \left( \frac{1}{2} \sum_{t=1}^{N} \left( y_{t+h} - \beta_0^h - \sum_{j=1}^{P} \beta_j^h x_{t,j} \right)^2 + \lambda \sum_{j=1}^{P} \left| \beta_j^h \right| \right) \tag{2.3}$$

| | |
|---|---|
| $(\hat{\beta_0^h}, \hat{\beta^h})$ | Estimation of the regression coefficients, |
| $\lambda$ | Hyper-parameter that determines the amount of shrinkage in the LASSO, |
| $(N, P)$ | Number of observations and variables. |

The implementation of the AR and LASSO models is performed with the glmnet package [87].

### 2.2.4 Nonlinear model

Being a linear model, the latter is not able to capture the wide range of weather behaviours. A more advanced model is thus needed. In the various scientific fields related to forecasting, two families of advanced models are strongly represented, namely Deep Neural Networks (DNN) and tree-based solutions. In [88], the authors compare both approaches in the light of top-ranked models in forecasting competitions as well as their distribution within the academic literature. It has been observed that tree-based models are often among the top contestants in forecasting competitions, while DNN receive more attention from the academic community. This phenomenon may be explained by the fact that the deep-learning field is currently excited about the possibilities of DNN and their numerous fields of application, and also because DNN are more prone to novel model work, which is a prerequisite for scientific publications. Nonetheless, the authors highlight that DNN models are often less robust than tree-based models: they require careful features scaling,

3. In this re-sampling procedure, the dataset is randomly partitioned into k equal-sized groups. Among these groups, one is retained as the test set, while the remainder k-1 groups serve as training data. A model is then fitted and evaluated. This process is repeated k times, in such a way that each sub-sample is used once as the validation data. The k scores can then be averaged to produce a single estimation.

hyper-parameter tuning, and skilful structural changes to obtain good performances on a novel task. In contrast, tree-based models can provide very good performances with default hyper-parameters. In addition, DNN are often more time-consuming to train and do not support feature importance calculation, which is beneficial to identify directions for improvement. In view of these elements, we decided to focus on tree-based algorithms because it is easier to tune the hyper-parameters of these models than to design the different layers of a deep neural architecture.

Forest-based models such as Random Forest (RF) or Gradient Boosted Regression Trees (GBRT) have been under active investigation for the last twenty years due to their good performances. Regarding this, the top models employed in the field of Renewable Energy Sources (RES) forecasting today are ensemble methods [4]. The recent forecasting competition organised by the 17$^{\text{th}}$ international conference on the European Energy Market (EEM20) was won by an architecture based on a Quantile Random Forest (QRF) model [89]. RF is often used as an advanced reference model to compare forecasting approaches due to its higher forecasting skills (e.g. its tolerance to poor information [90] and its tendency to give unbiased results [66]). Ultimately, the GBRT model has been ruled out due to the higher complexity of its hyper-parameter tuning. Generally, three main hyper-parameters have to be tuned (learning rate, depth of tree, number of trees) against one for the RF (number of trees), which generates longer training time. Moreover, contrary to models with higher complexity, RF can provide very good results without hyper-parameter tuning.

#### 2.2.4.1 Random forest model

RF [91] is a data-driven model able to perform nonlinear mapping between a set of input and output features. It is an ensemble learning method composed of several decision or regression trees grown in parallel.

First and foremost, let us focus on the constituents of RF models, namely: trees. The main idea behind tree models is to segment the inputs space into smaller coherent groups. To this end, a set of splitting rules is used at each node of the tree. To get a better glimpse of the tree algorithm principle, [92] proposes the following definition: "to build a prediction, trees ask each observation a series of questions, each one being in the form 'Is variable $X_j$ larger than a threshold s?' where j,s are to be determined by the algorithm" (Figure 2.3). As off-the-shelf models, decision/regression tree models have the advantage of being simple to implement, flexible and easily interpretable, but they also have a tendency to over-fit the training set (i.e. the model has a low bias but a high variance).

4. In statistics, ensemble methods are techniques based on a combination of multiple models with the aim of obtaining better performances than what could be obtained from the constituent algorithms.

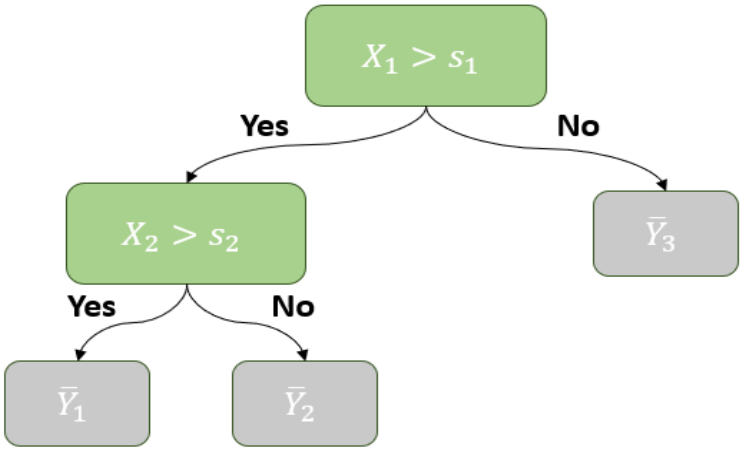



Figure 2.3 – Schematic diagram of the regression tree. $X_1$ and $X_2$ are two explanatory features while $s_1$ and $s_2$ are splitting criteria determined by the algorithm. $\hat{Y}_1$, $\hat{Y}_2$, and $\hat{Y}_3$ are the averaged values of the data in the terminal node.

One strategy to overcome the over-fitting issue and the lack of accuracy of the regression tree model is to combine multiple deep trees (i.e. to create a forest). To avoid obtaining similar or too-correlated trees while reducing the variance of the model, a bootstrap aggregating (i.e. bagging) algorithm is implemented, i.e. a random sample selection with a replacement of the training set is performed to feed each tree. Yet, it is still possible to obtain a correlated forest because of the strong predictors that are present in most of the trees. To tackle this issue, trees are fitted with a random set of features at each node (feature bagging). Lastly, the outputs of all trees are averaged (Equation 2.4), which reduces the variance of the model.

$$\hat{Y}_{t+h|t} = \frac{1}{T}\sum_{j=1}^{T} f_j(X_t) \tag{2.4}$$

$\hat{Y}_{t+h|t}$ Estimation of the response variable,
$X_t$ Vector of explanatory features,
$f_j$ j$^{\text{th}}$ regression tree,
$T$ Number of regression trees (Here, $T = 100$).

#### 2.2.4.2 Feature selection

RF has several built-in approaches for feature selection. One of them is based on *impurity* (i.e. a criterion to evaluate the goodness of splits). In the case of decision trees, each node can be viewed as a condition indicating how the input values are split into two sets, each of which contains values of the dependent variable that are similar and different from the other set. The importance of the feature is related to how *pure* the sets are. For regression, the measure of *impurity* is the variance. Trees naturally rank features by how effectively they improve the purity of the node: nodes with the greatest decrease in impurity are located

at the top of the trees, while features associated with the lowest decrease in impurity are located at the roots. As a result, one can keep most important features by pruning trees below a certain node.

#### 2.2.4.3 Features importance

Contrary to tree models, the main drawback of an RF lies in its lack of interpretability owing to the fact that predictions result from the averaging of a large number of tree outputs. Several options are proposed in the literature to counter this flaw [93]: the Mean Decrease Impurity (MDI) sums up the gain in purity associated with all splits performed along a given feature, while the *permutation importance* shuffles entries of a specific feature in the test set and computes the difference between the error on the permuted test set and the original test set (the features that have the biggest impact on performance are the most important ones) [92]. In this work, we use the MDI option.

The RF model is implemented with the ranger package [94], which provides a fast implementation of RF.

### 2.2.5 Forecasting paradigm

A forecasting model dedicated to a specific look-ahead horizon, $h$, is run in a rolling manner over the whole evaluation set. As a result, if $P_{t+h}$ represents the forecast PV power for time $t+h$, $t$ is a variable parameter, while $h$ is constant. In this paradigm, there are as many predicted time series as considered leading time (e.g. one predicted time series is associated with the 1-hour ahead horizon, another one with the 2-hour ahead horizon, etc.). In other words, a model is dedicated to a specific horizon. This treats the relative importance of the last observations differently depending on the considered horizons. From this point, the collection of 12 models (associated with the 12 horizons studied) will be simply referred to as the AR or RF model. This approach should be distinguished from plain forecasting, which delivers a set of forecasts associated with a unique launching time $t$. This approach is fitted for day-ahead forecasting: $t$ is a fixed parameter (usually, models are run in the morning), while $h$ is free. Moreover, to capture the specific features of each PV plant in terms of production characteristics and local atmospheric conditions, a forecasting model is fitted for each site. In a nutshell, we adopt a single-site and single-horizon forecast architecture.

The models are trained over the year 2015 and evaluated on the period covering 2016. The input explanatory variables and the PV power forecast outputs have a 15-min granularity.

## 2.3 Evaluation concept

To assess a forecast, it is necessary to first answer the question *what is a good forecast?* From an economic point of view, forecasts may be assessed by their influence over decision-makers, and the subsequent economic value that they may generate [35]. While, from a forecasting perspective, a good forecast may be considered as a forecast that produces small errors in light of observed data. In other words, do we measure the profitability of a forecast (in the sense that it can generated additional value in decision-making processes) or its accuracy?

In this document, we focus on the second aspect of forecasts, which leads to another question: how can we measure forecast accuracy? The literature provides a large range of metrics tailored for specific uses, for instance, Zhang et al. in [95] propose a suite of 16 metrics, while [96] introduces two new metrics of which one quantifies the ability of forecasts to follow ramp events. The verification framework developed in this section is based on a set of well-established scoring rules and on visual diagnostic tools used or encouraged by the literature. A recent paper [97], written in 2020 by many prominent researchers in the solar forecasting field, points out the lack of standardised methods to verify deterministic solar forecasts. As a response, the authors propose a general verification framework to facilitate forecast analysis and comparison within the literature. The main recommendation is to use two complementary approaches to assess forecast quality and to assist forecasters to make informed decisions; namely a measure-oriented approach and a distribution-oriented approach. The former, based on the Root Mean Square Error (RMSE) skill score, is recommended for cross-work forecast comparison and is a good indicator of the global skilfulness of the model. The second verification framework is based on the joint distribution of forecasts and observations. To give a better idea of these approaches, we can also classify them respectively as quantitative and qualitative metrics. The former condenses the performance into a single value, while the latter offers a visual representation of error distribution. In this regard, visualisation of forecast performances is an efficient way of communicating the performance of a model. Thus, this medium is preferred over tables of values.

At this point, we draw the reader's attention to the fact that performances are always evaluated based on predictions that integrate the Sun's path dependency. In the case of clear-sky normalised inputs, predictions are post-processed to reinstate the Sun's path dependency. This is done by multiplying the clear-sky normalised prediction with the associated value of the clear-sky model output.

### 2.3.1 Benchmark model: clear-sky index-based persistence model

In the literature on RES forecasting, the persistence approach is often used as a reference. In simple words, the latter supposes that the meteorological situation does not change over

time, consequently, prediction is equal to the last observation. This model only uses past measurements and does not involve any modelling process. The main assumption behind this approach is that the weather situation, and so the related PV generation, remains unchanged for a certain amount of time. Despite being a naive approach, the persistence model exhibits good performances for very short-term horizons for which the persistence in cloud structures and distribution can be observed, and for situations with low weather variability (e.g. clear-sky or very cloudy days). On days that are entirely clear or entirely overcast, predictions of PV generation are straightforward: the CSI may be assumed as constant (e.g. 1 in case of clear-sky days). In these specific cases, the persistence model performs pretty well, and its forecasts can be difficult to surpass, especially for the very first time-steps. On the contrary, on days with partially or intermittently cloudy skies, PV generation forecasting becomes more challenging, and the persistence model exhibits comparatively poor performances. As a result, it is often used as a fallback model when advanced approaches fail.

In the literature, the persistence model features different formulations. For instance, the forecast output for time $t+h$ can either be equal to the last observation at time $t$ [46], to the observation on the previous day at the same leading time $t+h$ [98], or even to the mean of the $h$ previous observations [55]. Here we consider a clear-sky index-based persistence model [97], which takes into account the yearly and daily seasonal cycles. Generally, persistence-based models are blind when forecasts are generated at night, which artificially restrains the performance of early morning forecasts. This promotes the generation of under-optimistic forecasts, and consequently over-optimistic comparisons with other models. To tackle this issue, we propose to consider the observations of the previous day at the same leading time $t+h$ for forecasts issued at night. This leads to the model defined in Equation 2.5. Thereafter, this model is simply denoted as *persistence*.

$$\widehat{\overline{P}}_{t+h|t} = \begin{cases} \overline{P}_t & \text{if } P_t \neq 0 \text{ (i.e. daytime)} \\ \overline{P}_{t+h-24H} & \text{if } P_t = 0 \text{ (i.e. nighttime)} \end{cases} \tag{2.5}$$

$\bar{.}$ The quantity is normalised by a clear-sky-based feature,
$\overline{P}_t$ Clear-sky based normalised production observed at time $t$,
$\widehat{\overline{P}}_{t+h|t}$ Expected PV production at time $t+h$ based on elements available at time $t$.

### 2.3.2 Quantitative forecasting performance criteria

A large range of performance metrics has been defined by the scientific community, each of which highlights a specific aspect of the forecasting error [96]. In the present study, we use a set of well-established metrics to characterise the accuracy of the forecasting models, while enabling comparison with other studies. In this respect, PV production is normalised by the corresponding installed capacity of the site, $P_c^x$, to prevent dependence on power plant size.

The forecasting process is performed with clear-sky normalised inputs (the normalisation process is detailed in Section 4.4.1). To compare production observations with forecasts, it is necessary to obtain de-normalised outputs. To do so, the forecast outputs are multiplied with the associated clear-sky model outputs. In addition, nighttime data are discarded from the validation framework inasmuch as they do not offer relevant information. Observations associated with low-Sun situations (i.e. zenith angles higher than 85°) are excluded from the whole forecasting framework because the clear-sky based normalisation process has some weaknesses (to be precise, the CSI reach unrealistic values) when used at early and late hours of the day. Rather than implementing complex correction strategies like in [58], we opt to reject fallacious values because the irradiance is too low to be significant in solar power applications [97].

It is futile to base our quality analysis on a unique accuracy score, because two drastically different forecasts can lead to the same scores [96]. Thus, we consider the three most popular metrics [80], namely the capacity-normalized Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE), which are respectively described by Equations 2.6, 2.7, and 2.8. These metrics provide information on the long-term performance of a model. The normalised Root Mean Square Error (nRMSE) score is based on the square of the forecast error, while the normalised Mean Absolute Error (nMAE) considers the absolute value of the forecast errors' amplitude. Thus, the main difference between the nRMSE and nMAE is that the former is very sensitive to large errors and outliers, while the latter gives the same weight to all errors. The MBE describes the unconditional bias. A positive/negative normalised Mean Bias Error (nMBE) represents an over-prediction/under-prediction, where on average forecasts are higher/lower than observations. As pointed out in [97], a small MBE is more of a baseline requirement rather than a creditworthy feature among state-of-the-art forecasts because of the possibility to correct bias thanks to Model Output Statistics (MOS) approaches. One drawback of this score is that over-estimation and under-estimation may cancel each other out in separate observations. It is important to note that throughout this work, model parameters are estimated with a loss function analogous to the RMSE. As such, it is not surprising to observe forecasts that excel more according to the nRMSE criterion than to the nMAE criterion.

$$nRMSE^{x}(h) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} - \boldsymbol{P}^{\boldsymbol{x}}_{\boldsymbol{t+h}}}{P^{x}_{c}}\right)^{2}} \tag{2.6}$$

$$nMAE^{x}(h) = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} - \boldsymbol{P}^{\boldsymbol{x}}_{\boldsymbol{t+h}}}{P^{x}_{c}}\right| \tag{2.7}$$

$$nMBE^{x}(h) = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} - \boldsymbol{P}^{\boldsymbol{x}}_{\boldsymbol{t+h}}}{P^{x}_{c}}\right) \tag{2.8}$$

| | |
|---|---|
| $N$ | Number of paired data, |
| $P_c^x$ | Installed capacity of site $x$, |
| $P_{t+h}^x$ | Observed production at time $t+h$ and site $x$. |

On the one hand, due to its capacity to penalise large errors more heavily, the nRMSE score seems appropriate to meet system operators' concerns, as large imbalances require committing more costly reserves. On the other hand, the nMAE score appears as an interesting indicator for energy producers inasmuch as financial penalties are usually proportional to the absolute imbalance between the forecast and the actual production. Consequently, these three error metrics are displayed in parallel in all evaluation steps.

Scores are computed individually for the nine PV farms, but we average them for a more compact presentation. Thus, we obtain scores that are only horizon- and model-dependent.

To gauge the skilfulness of a forecasting method, it is relevant to compare it with a reference model that can sufficiently reflect the difficulty (variability and uncertainty) inherent to a forecast situation [97]. The persistence model is a suitable candidate for this goal. Such a comparison offers a common basis with the literature for comparison purposes. We consider the skill score defined as follows:

$$SS_M(h) = \frac{A_M(h) - A_{Ref}(h)}{A_P(h) - A_{Ref}(h)}. \tag{2.9}$$

| | |
|---|---|
| $A$ | Measure of accuracy (e.g. nRMSE or nMAE), |
| $A_M(h)$ | Accuracy score obtained with model $M$ for horizon $h$, |
| $A_{Ref}(h)$ | Accuracy score of the *Persistence* model for horizon $h$, |
| $A_P(h)$ | Accuracy score of a perfect forecast (for the nRMSE and nMAE metrics, a perfect forecast implies $A_P(h) = 0$). |

The skill score is expressed as a percentage, representing the relative accuracy improvement of the studied model over the reference model. A positive (negative) skill score implies that the forecasting model has a smaller (higher) score than the reference method. A skill score equal to zero means that the performances of both models are equal, while a perfect forecast is obtained for a skill score of one. For instance, the skill score based on the nRMSE is defined as:

$$SS_M(h) = 1 - \frac{nRMSE_M(h)}{nRMSE_{Ref}(h)}. \tag{2.10}$$

As the nMBE of the persistence method is often close to zero, the associated skill score is undefined, which makes the nMBE score unsuitable for skill score computation considering the persistence as the reference model.

### 2.3.3 Significance of forecast differences

Throughout this thesis, we have been confronted with forecasts issued by different models but characterised by very close accuracy scores. In this context, it becomes challenging to determine whether the differences are statistically significant.

To assist us in this process, we turn to the Diebold-Mariano (DM) test. The DM test compares the predictive accuracy of two forecast models. The time loss differential between the two forecasts is denoted by $d_{12,t} = g(e_{1,t}) - g(e_{2,t})$ with $e_{i,t}$ being the forecast error, and $g$ an arbitrary loss function. The two forecasts have equal accuracy if the expectation of the loss differential is zero (which constitutes the null hypothesis: $H_0 : E(d_{12,t}) = 0, \forall t$). Under the null hypothesis and for large samples, the DM test follows the standard normal distribution (Equation (2.11)) [99]. We assume a significance level of 5%. As a result, DM statistics that fall outside the range defined by the 2.5% and 97.5% quantiles of the normal distribution (i.e. $-1.96$ and $+1.96$) enable the rejection of the null hypothesis.

$$DM_{12} = \frac{\overline{d}_{12}}{\hat{\sigma}_{\overline{d}_{12}}} \sim \mathcal{N}(0,1) \tag{2.11}$$

$\overline{d}_{12}$ The sample mean of the loss differential series ($\overline{d}_{12} = \sum_{t=1}^{T} d_{12,t}$),
$\hat{\sigma}_{\overline{d}_{12}}$ A consistent estimate of the standard deviation of $\overline{d}_{12}$ ([99]).

### 2.3.4 Qualitative forecasting performance criteria

Following the recommendations of [97], we implement the Murphy–Winkler framework for distribution-oriented forecast verification. Within the scope of this work, this framework takes the form of a forecast–observation scatter plot containing information regarding joint and marginal distributions. Since it contains all of the time-independent information about the forecast performance, this framework provides more information than the measure-oriented method introduced in the previous section. Interested readers may refer to the above-mentioned article regarding mathematical background of this framework. Illustrations are provided in Figures 2.13 and 2.14.

Although this diagram provides a clear visual tool to inspect the quality of point forecasts, it is not suitable to efficiently compare of the different forecasting architectures investigated throughout this work. Instead, the measure-oriented framework is used for performance comparison between models, while this distribution-oriented framework is used for performance assessment of the best forecasting architectures developed.

## 2.4 Data overview

This section focuses on the different sources of information considered for PVPF. Proven data consist of All-Sky Imagers (ASI), production and satellite-derived information, and NWPs model outputs (Figure 1.8). In the literature [55, 60, 100], we observe a shift towards

models combining multi-source inputs due to their capacity to improve forecasting accuracy. This explains why the specialised literature is still growing and new sources of information are currently being investigated; for instance, [53] experimentally demonstrates the use of a network of ASI dedicated to PVPF nowcasting. The Smart4RES project aims at, inter alia, developing a collaborative RES forecasting approach by providing a framework for data sharing that preserves confidentiality constraints, and an incentive data market [60].

In addition, the pool of investigated data sources has shifted from temporal-based sources of information to Spatio-temporal (ST)-based inputs. In the present context, the notion of *spatio-temporal* data may be understood as a set of physical quantities that have the same dimension, and are measured or computed at several spatial points. Such features are considered together in the forecasting architecture with the aim of valuing spatial and temporal dependencies on PV production. The latter has the advantage of providing observations of the spatial distribution of solar irradiance, and giving a glimpse of forthcoming cloud structures with spatial and temporal resolutions depending on the nature of the sensor. On the other hand, *temporal* data refer to the physical parameters considered at the power plant location, which are employed due to their temporal dependencies on power production.

In the scope of this work, production measurements, satellite-based observations and NWPs model outputs have been considered for their proven interest regarding PVPF (Figure 2.4). With the aim of extending the current portfolio of available data, we could have envisaged adding (1) non-professional weather station network observations, or (2) on-site weather observations. The former option would be interesting owing to the democratisation of connected personal weather stations, which provides varied and dispersed measurements of physical parameters. Similarly, on-site observations of atmospheric parameters such as wind velocity and direction could have been used as part of an ST-based approach [62]. Nevertheless, these options were not further investigated due to issues regarding missing data management, quality control, and database rights.

We choose to feed short-term forecasting models with data that have a 15-min temporal resolution. This time-step makes it possible to obtain a fine vision of weather variability and is adapted to grid balancing. PV production and satellite observations are provided at this time-step, but the NWPs model outputs still require a temporal interpolation.

### 2.4.1 PV production observations

We consider production records from nine fixed-tilt PV grid-connected systems located in the south-eastern part of France, mainly along the Rhône river (Figure 2.5) and operated by the Compagnie Nationale du Rhône (CNR).

Despite all sites being located in the same region, climate conditions vary from one place to another: the southernmost sites are mainly influenced by a Mediterranean climate while the westernmost site is subject to an Alpine climate. To get an insight into cloud structure dynamics, we adopt the days classification proposed in [55] based on intraday clear-sky

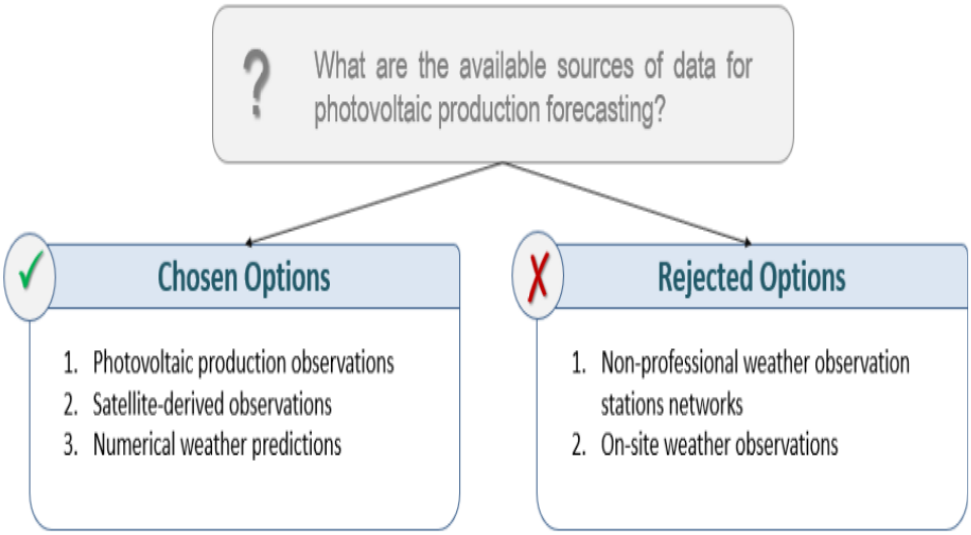



Figure 2.4 – Available options regarding data sources.

| | **PV1** | **PV2** | **PV3** | **PV4** | **PV5** | **PV6** | **PV7** | **PV8** | **PV10** |
|---|---|---|---|---|---|---|---|---|---|
| **PV1** | 0.00 | 42.90 | 27.43 | 94.74 | 48.34 | 41.64 | 60.55 | 53.67 | 98.65 |
| **PV2** | | 0.00 | 26.06 | 80.36 | 42.64 | 41.79 | 53.82 | 82.96 | 133.05 |
| **PV3** | | | 0.00 | 100.08 | 55.97 | 51.75 | 68.80 | 56.96 | 125.87 |
| **PV4** | | | | 0.00 | 46.48 | 53.13 | 34.32 | 148.31 | 122.85 |
| **PV5** | | | | | 0.00 | 7.35 | 12.93 | 101.84 | 102.90 |
| **PV6** | | | | | | 0.00 | 18.92 | 95.29 | 98.21 |
| **PV7** | | | | | | | 0.00 | 114.20 | 103.08 |
| **PV8** | | | | | | | | 0.00 | 124.32 |
| **PV10** | | | | | | | | | 0.00 |

Table 2.1 – Distance between pairs of sites (in km).

average, $k_\mu^3$, and Intraday Variability (IV). On the one hand, the intraday clear-sky average (Equation 2.12, with $k_t^3 = \frac{PV_t}{PV_t^{sim}}$, the CSI defined in Section 4.4.1.2) enables us to assess the day's weather type, from overcast to clear atmosphere. On the other hand, the IV (Equation 2.13) accounts for the weather variability of the day, from stable to very variable throughout the day.

$$k_\mu^3 = \frac{\sum_{t=1}^{N_d} k_t^3}{N_d} \tag{2.12}$$

$$IV_t = \sqrt{\frac{\sum_{t=1}^{N_d} \left(\Delta k_t^3 - \Delta k_\mu^3\right)}{N_d}}, \text{ with: } \Delta k_t^3 = k_{t+1}^3 - k_t^3 \tag{2.13}$$

$N_d$ Number of observations of the day (except nighttime data).

The resulting classification is depicted in Figure 2.6. The southernmost power plants experience a higher rate of stable sunny days (type CI) during summer, and are characterised

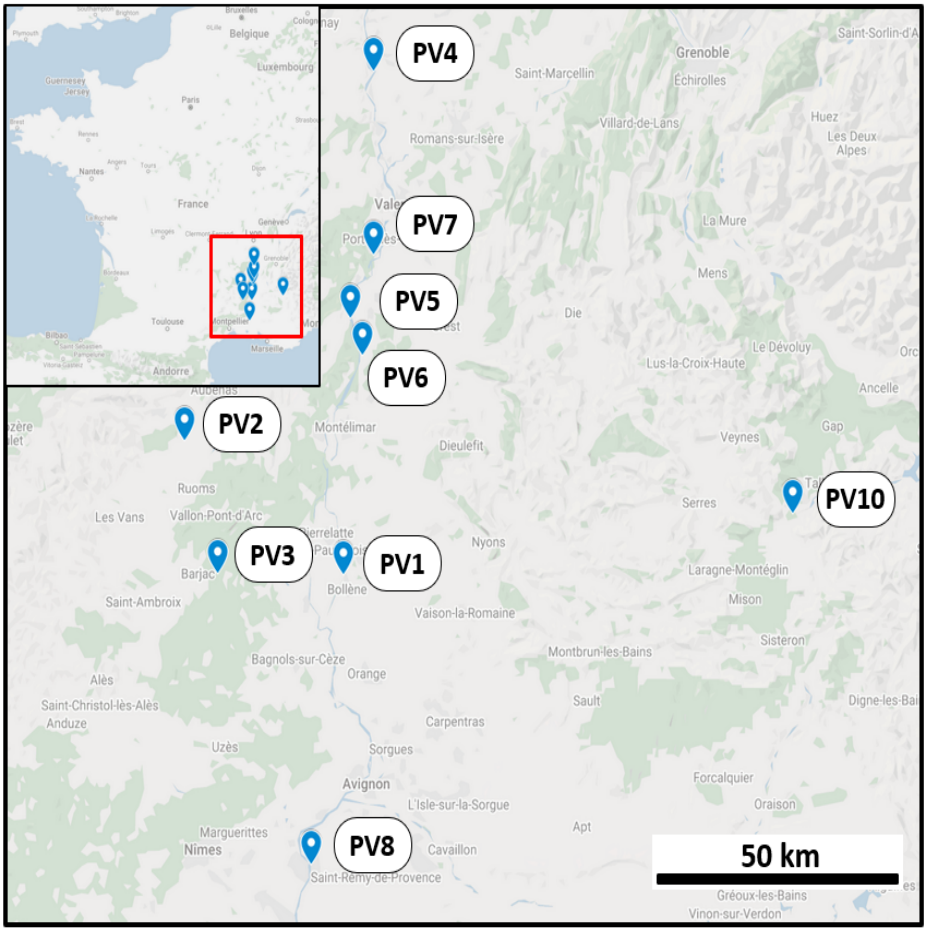
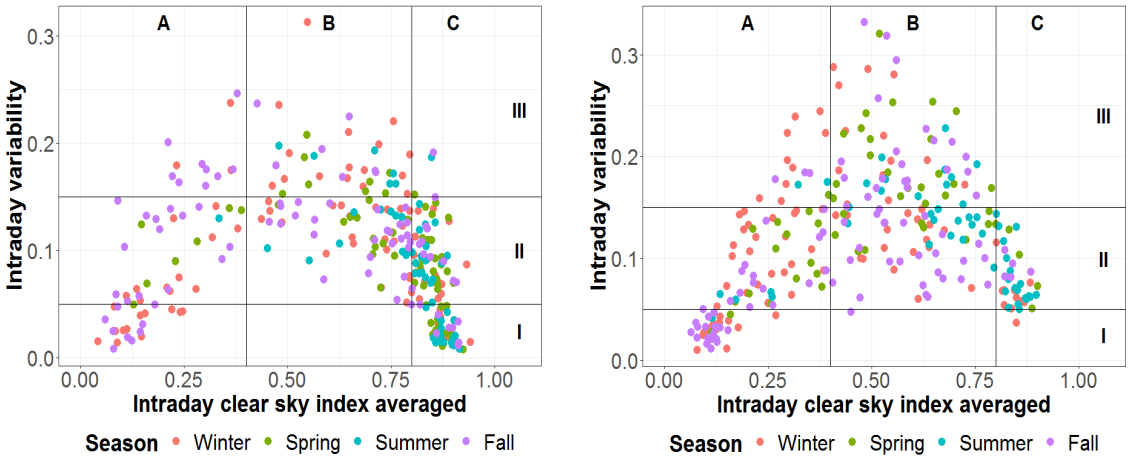



Figure 2.5 – Spatial distribution of CNR's PV sites, represented by the blue spots. These power plants are located in southeast France.

by a dichotomous classification (globally such sites experience either overcast or sunny days). Type B situations are highly variable, which suggests daily situations alternating between sunny and cloudy weather. Northernmost sites experience more days with scattered clouds (types BIII and BII), and are characterised by stable, overcast autumn days (type AI). Comparatively, PV4 have fewer clear-sky days (type C) than the southernmost sites.

Due to its geographical position, PV10 is the only power plant subject to an Alpine climate, and as such, it is more likely to experience low production rates resulting from snow deposition. Snow deposition on modules may induce bias during model learning because low production may not be associated with low incoming irradiance. Thus, snowfall from the ERA5 reanalysis dataset is used to identify and reject days with accumulated snow greater than 3 cm and an associated PV production lower than 10% of installed capacity (which corresponds to 12 and 19 days for the training and testing periods respectively).

The installed power capacity ranges from 1.2 to 12 MWp, occupied areas vary from 1.3 to 12.0 ha (Table 2.2), and the distance between sites ranges from 7.3 to 133 km (Table 2.1). To allow comparison between generation units, PV production is normalised by the corresponding installed capacity $P_c$. Regarding this, power output is limited by the downsizing of




(a) Weather variability observed on PV8 (southernmost site).

(b) Weather variability observed on PV4 (northernmost site).

Figure 2.6 – Distribution of days according to their intraday variability and intraday CSI (described in Section 4.4.1) average adapted from [55]. High values of CSI are rejected. Roman numbers (I, II, III) classify the variability, from stable to very variable weather. Letters (A, B, C) represent weather type from overcast days to clear days.

the inverters. Such an approach is part of a trade-off between maximising energy yield and minimising inverter costs [101]. Large power plants are subject to spatially variant ambient environments [102]: for a given time, PV strings located farther away may exhibit distinct power profiles due to being exposed to various irradiance levels. Such ST characteristics can be valuable for nowcasting applications. Intuitively, the larger the plant, the lower the power fluctuations. Yet, the standard deviation between PV2 and PV3, which are close enough to experience the same weather climate, is very similar. This may result from an excessively coarse sampling time [103].

| **Site Name** | **Installed power (MWp)** | **Occupied area (ha)** | **Orientation angle (°)** | **Tilt angle (°)** |
|---|---|---|---|---|
| PV1 | 4.05 | 2.85 | 176 | 25 |
| PV2 | 12 | 7.65 | 180 | 25 |
| PV3 | 1.3 | 0.82 | 180 | 25 |
| PV4 | 4.21 | 2.65 | 180 | 25 |
| PV5 | 3.43 | 2.25 | 180 | 25 |
| PV6 | 4.12 | 3.13 | 180 | 25 |
| PV7 | 2.42 | 1.53 | 180 | 25 |
| PV8 | 2.94 | 2.39 | 180 | 25 |
| PV10 | 2.88 | 1.74 | 180 | 25 |

Table 2.2 – Technical configuration of PV sites.

To build an effective model, it is essential that the model be trained and tested on

data that share similar statistics. As a reminder, data from 2015/2016 are used during the training/testing step. For instance, Figure 2.7 shows the normalised variability defined by Equation 2.14 for each PV unit for the training and testing datasets (the CSI, $k^3$, is defined in Section 4.4.1.2). Overall, the testing set contains slightly more variability than the training set: the average variability of the testing sets is around 4% higher than the average variability of the training dataset. However, we assume that this low difference is enough to consider that both datasets are compatible.

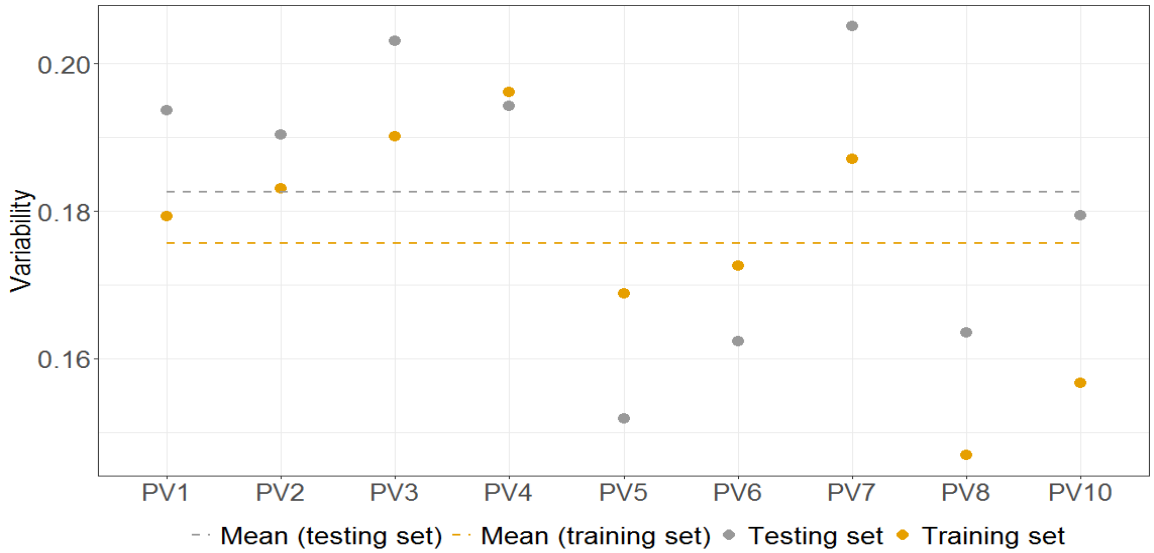


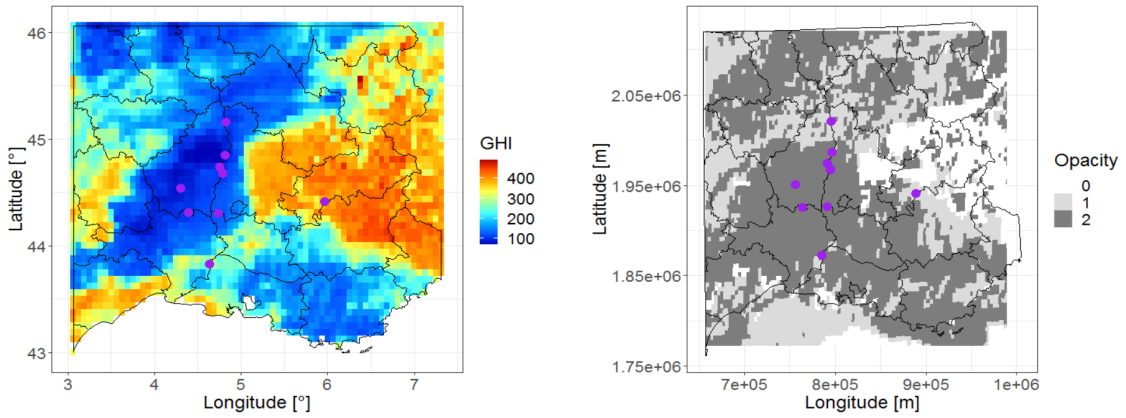

Figure 2.7 – Normalised variability of each PV site for the training and testing datasets. Dashed lines represent average values over the nine PV sites.

$$V = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(k_t^3 - k_{t-1}^3)^2} \tag{2.14}$$

### 2.4.2 Satellite-derived data

Two sources of satellite-derived information are investigated. The most widespread one is the Satellite Derived Surface Irradiance (SDSI) (Figure 2.8a), which represents estimations of solar irradiance reaching the ground. When forecasts are generated during the night for the early morning, no relevant solar or past production observations are available, which obliges the forecasting model to propose average values learnt during the training phase. That is why we also consider cloud opacity maps derived from cloud classifications, which are themselves generated from the infrared channels of satellites. Such datasets enable us to access the clouds' position during the nighttime (Figure 2.8b). To the authors' knowledge, this kind of input is still marginally used in the literature (e.g. [104] considers satellite infrared images and the forecast of wind velocity to propose a cloud motion forecasting method for the morning). Nevertheless, the state-of-the-art regarding the use of visible satellite-based information highlights its relevance for horizons up to 6 hours ahead, which

may suggest similar benefits for infrared-based information.




(a) SDSI.

(b) Cloud opacity (0 meaning a lack of clouds).

Figure 2.8 – Satellite-based maps of estimated GHI on the left, and cloud opacity on the right observed on 2015-01-15 11:00:00. Purple points show the position of PV units.

Both sources of information are obtained with the geostationary Meteosat satellites. Satellite-derived maps possess a spatial resolution of 3 km at the nadir and a temporal resolution of 15-min. Depending on the forecasting tools used, satellite observations may be considered either as a sequence of 2D maps or as a set of time series derived from each pixel constituting these maps.

#### 2.4.2.1 Ground irradiance

The SDSI data are extracted from the Helioclim-3 database [105], which stores 15-min GHI maps with around a 5-km spatial resolution in Europe ($0.0625° \times 0.0625°$). This database is generated by the Heliosat-2 method [106], which processes images collected by meteorological geostationary satellites into maps of solar radiation. In simple terms, this conversion process is performed by combining the output from clear-sky models (i.e. estimation of the ground irradiance considering a cloudless sky) with a transmittance function representing the impact of all cloud layers and surface interactions on the solar irradiance at the Earth's surface [107].

#### 2.4.2.2 Cloud opacity

Opacity maps are derived from cloud classification maps provided by Meteo-France. Classification maps are generated by the geostationary Meteosat second-generation satellites. These satellites observe the Earth in 12 spectral channels, of which 8 are dedicated to thermal infrared [108]. Each of these channels is associated with specific properties of atmospheric air masses (e.g. the 8.7 $\mu m$ channel provides quantitative information of thin cirrus clouds) [109], which allows us (1) to generate a cloud mask showing the location of clouds, and then (2) to allocate a meteorological cloud type to the identified clouds thanks

to threshold-based algorithms [110, 111].

The clouds are classified into 19 types and associated with 3 levels of opacity (Table 2.3). Cloud classifications in fact provide more detailed information but it is more difficult to process. In accordance with internal practices performed at CNR, we choose to work with opacity coefficients.

| Code | Designation | Opacity coefficient |
|---|---|---|
| 0, 20 | Not coded | |
| 1-4 | Cloudless pixel | 0 |
| 6, 7 | Very low clouds | 2 |
| 8, 9 | Low clouds | 2 |
| 10, 11 | Middle clouds | 2 |
| 12, 13 | High opaque clouds | 2 |
| 14 | Very high opaque clouds | 2 |
| 15-17 | High translucent clouds | 1 |
| 18 | High translucent clouds above low or middle clouds | 2 |
| 19 | Scattered clouds | 1 |

Table 2.3 – Coding used in cloud type classification and associated opacity coefficients.

In our case study, as data are only available from 04:00:00 UTC (Figure 2.9), this input is mainly relevant for wintertime forecasts when the sunrise occurs a few hours later, but has a very limited contribution during summertime when the Sun rises early in the morning.

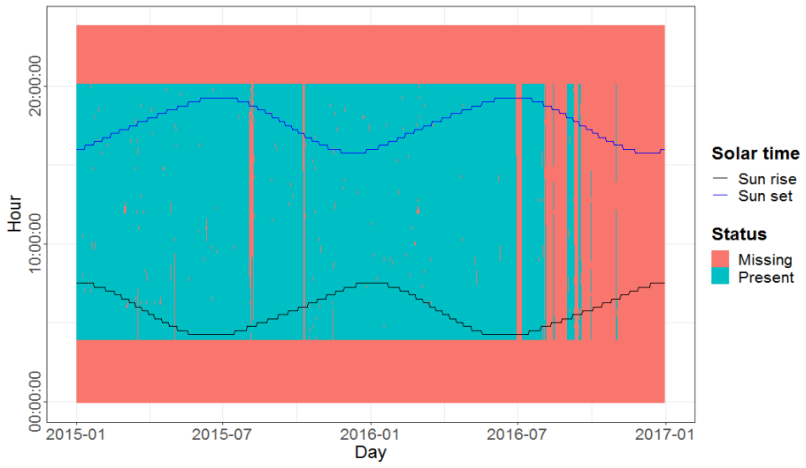



Figure 2.9 – Opacity maps availability.

### 2.4.3 Numerical weather predictions

The NWPs used in this work are obtained from the highest resolution (HRES) configuration of the Integrated Forecast System (IFS) run by the European Centre for Medium-

Range Weather Forecasts (ECMWF). This model is run twice a day, at 00:00:00 UTC, and 12:00:00 UTC providing parameters with a 1-hour temporal resolution and a $0.1° \times 0.1°$ spatial resolution.

These discretisation scales result from a trade-off with the computational cost (smaller scales involve determining more parameters). Therefore, sub-grid-scale weather phenomena such as small cloud generation cannot be explicitly determined by solving the model's physical equations. Instead, a parameterisation procedure based on physical representation (e.g. radiation law) or statistical laws (e.g. inferring cloudiness from relative humidity) is used to approximate small/brief, complex or poorly understood processes. The main limitations of NWPs are their coarse spatial and temporal resolution: the former makes it impossible to resolve most clouds but only provides an average cloud cover, while the latter does not enable the assessment of time-dependent cloud cover variability [46]. Nevertheless, NWPs provide valuable information regarding weather trends.

#### 2.4.3.1 Use of the NWPs model

NWPs models are computed several times a day (here at 00:00:00 UTC and 12:00:00 UTC). These sets of forecasts are named *runs*. In an operational context, a run may need up to 6 hours of computational and data delivery time before being available for end-users. In the present paper, we neglect this aspect and consider that forecasts are available at the launching time of the run.

Depending on the lead time, several predictions can be issued for the same time (e.g. predictions for time 13:00:00 can be provided by the runs of 00:00:00 and 12:00:00 on the same day). As a result, two approaches are considered according to the weather information integration strategy.

First, one may consider that each run has distinctive features: the number and position of initial observations used to initialise the numerical model may vary according to its launching time, which may impact the quality of the forecasts. In other words, for the same lead time, two runs may have different forecasting precision and bias. Therefore, when NWPs are considered as state features, it is relevant to compare predictions with similar errors. To do so, we consider runs delivered at the same time of day to characterise weather situations (e.g. if the predictors describing the target situation for time $t + h$ come from a 12:00:00 run, then the predictors describing the candidate situations also come from 12:00:00 runs).

Alternatively, one may focus on the fact that forecasting precision tends to decrease as the lead time increases. As a result, when NWPs are considered as explanatory variables, only predictions from the most recent runs are considered.

#### 2.4.3.2 Local characterisation of the atmosphere

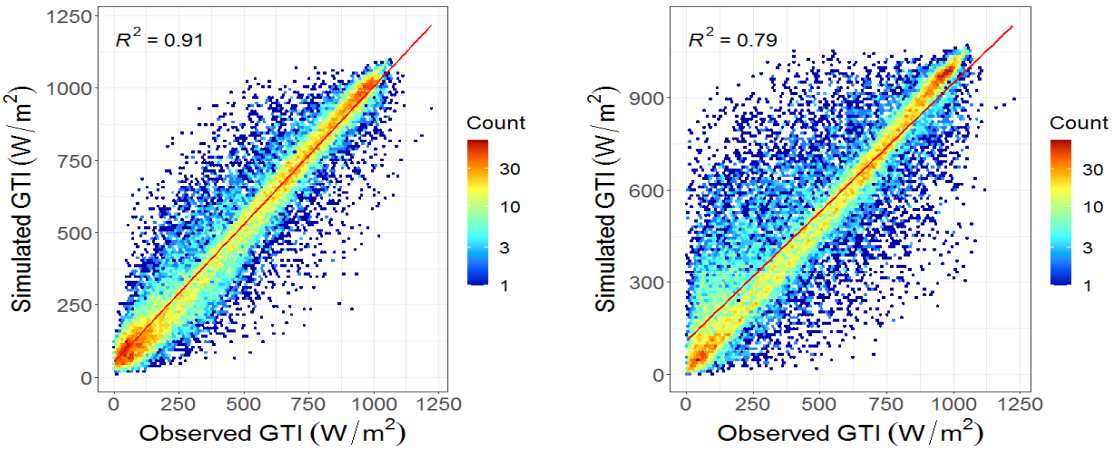
**2.4.3.2.1 Choice of parameters** In the PVPF field it is common practice to resort to surface sensible weather features to account for PV production (e.g. GHI, 2-m Temperature (T2M), and 10-m wind). Forecast parameters such as T2M, and Total Cloud Cover (TCC) are computed by the physical parametrisation part of the IFS model (i.e. an approach which replaces processes that are too small-scale or complex to be modelled physically by simplified expressions), while irradiance is computed by a radiative transfer model based on predicted values of temperature, humidity, cloud, and monthly mean aerosol climatologies [112]. Here, we consider the following parameters: GHI, TCC, T2M for their proven interest in PVPF [113]. These parameters are considered at the site position through a bi-linear interpolation of the nearest grid points.

**2.4.3.2.2 Modelling errors of irradiance features** To get an idea of the errors contained within features characterising the local atmospheric states (i.e. satellite-based observations and NWPs), the latter are compared with on-site observations of irradiance. On the one hand, irradiance forecasts are obtained by gathering forecasts from the latest runs. This alleviates forecast errors, which tend to grow as the forecast horizon extends. On the other hand, we focus on satellite-based information derived at the site location.

As our datasets of on-site observations lack GHI-based features, we turn to Global Tilt Irradiance (GTI) [5] measured by reference cells. This compels us to project the GHI, provided by satellite observations and the NWPs model, into the Plane-of-Array (POA). This projection model is detailed in Section 3.3.1.2.

Figure 2.10 highlights that both estimated GTI derived from (1) satellite-based GHI, and (2) GHI predictions issued by the numerical weather model are centred on the identity line, and match well with the vast majority of corresponding measurements. Yet, in both figures, we observe an offset between the lowest values of simulated and measured irradiance. This is thought to result from shading effects altering irradiance reception for low elevation angles. In general, the dispersion of the scatter points results from mismatches between simulated and observed values. These mismatches may be due to the coarse spatial resolution of data (e.g. predicted data cannot assess small cloud structures which directly affect PV plant production). In addition, differences may be accounted for by a mismatch between the atmosphere turbidity at the site location and the climatologic values used in the Helioclim and IFS models, and also by errors induced by the projection model. A higher dispersion of the scatter points is observed in Figure 2.10b. This results from the facts that (1) 15-min observations are compared with 1-hour based interpolated predictions, (2) the spatial resolution of the NWPs is coarser than that of the SDSI, and (3) uncertainties are present in predictions.

5. GTI represents the solar radiation incident on a tilt surface (in our case, it corresponds to the tilt angle of the PV panels).




(a) Modelled GTI obtained from SDSI.

(b) Modelled GTI obtained from NWPs considering latest forecasts of each run.

Figure 2.10 – Binned scatter plots for observed and modelled GTI of PV1 during 2015 and 2016. GHI is projected into GTI thanks to the conversion model detailed in Chapter 3. GTI is measured by reference cells. The red line represents the regression line obtained with the least-squares method.

#### 2.4.3.3 Global characterisation of the atmosphere

Sensible weather occurs on small scales and strongly depends on larger-scale features. NWPs models have some difficulty forecasting phenomena whose the governing processes occur at sub-grid scales, like explicit cloud formation, but turn out to be much more reliable at forecasting large-scale atmospheric fields, such as synoptic-scale pressure fields, insofar as such parameters are explicitly resolved within the models.

In the meteorological forecasting domain, geopotential fields (i.e. representation of large-scale pressure patterns in the atmosphere) are commonly used to forecast precipitation generation [68] and demonstrate strong influence over wind direction. They represent the geopotential height at which the corresponding atmospheric pressure level is reached (e.g. considering a 925 hPa pressure level, if at a specific location the geopotential height is 5,300 m, it means that a 925 hPa atmospheric pressure is achieved at 5,300 m above sea level). From the geopotential fields one can derive the pressure gradient that drives the air flow from high to low pressure regions, namely the geostrophic wind. In summary, geopotential height is highly correlated with air flow and cloud generation, which makes it suitable to work with PVPF.

## 2.5 Preliminary results

This section introduces preliminary results regarding forecasts generated solely from past PV production observations.

Figure 2.11 gives a visual representation of the measure-oriented framework developed

in Section 2.3.2. The left panel is composed of the nRMSE, nMAE, and nMBE scores, while the right panel displays associated skill scores with respect to the persistence model. Due to the very low bias of the persistence, the nMBE-based skill score contains limited information. We observe that the three models exhibit very low bias. The persistence model tends to under-estimate production, while the AR and RF models provide over-estimated forecasts. In terms of nRMSE, both the AR and RF models outperform the persistence approach. This may be partially accounted for by the fact that the parameters estimation of the AR model is based on the Mean Square Error (MSE) loss function. On the contrary, the nMAE scores of the AR model are lower than those reached by the persistence. The RF model outperforms the AR model for both nRMSE and nMAE accuracy scores.

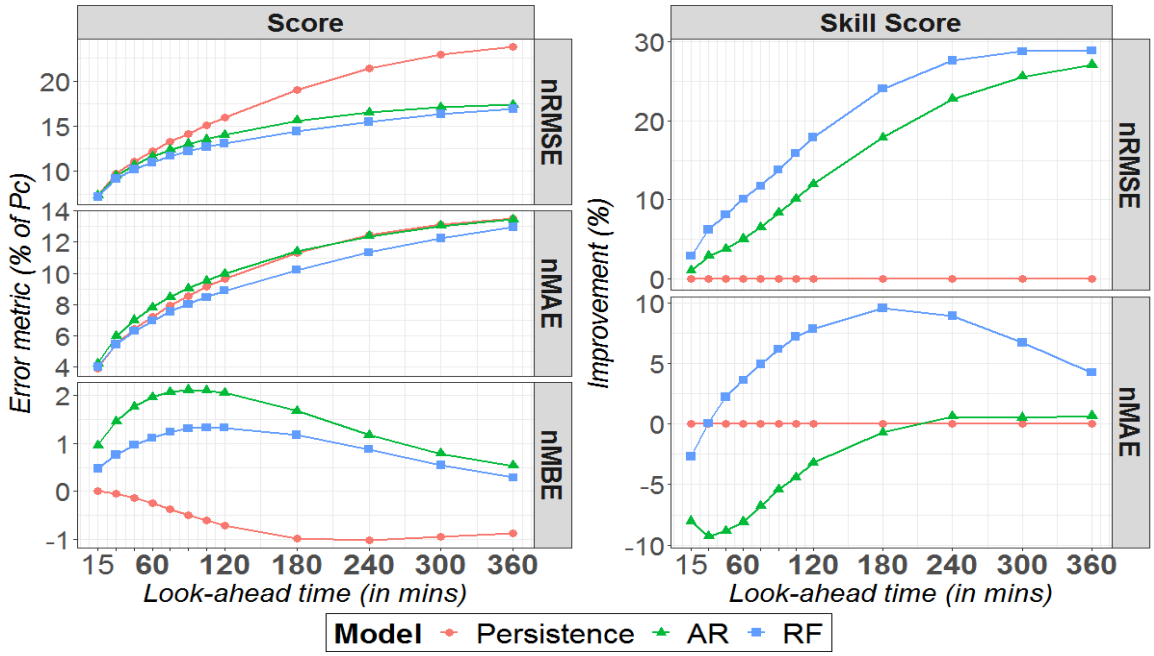



Figure 2.11 – Forecasting performances obtained considering former PV production observations. Data have been quality checked with the method presented in Section 4.2, and normalised with the method introduced in Section 4.4.1.

The nRMSE scores achieved within the scope of this study using the AR model are in line with what can be observed in the literature (Table 2.4). Any divergences are assumed to result from sites' specific features (e.g. plant size, local weather).

| Study | Location | 1-hour | 6-hour |
|---|---|---|---|
| [83] (Fig 7) | Portugal | 8.5% | 13.7% |
| [100] (Fig 2.15 (b)) | France | 6.25 – 11.5% | 7 – 18.75% |
| Current study | France | 11.8% | 17.8% |

Table 2.4 – Comparison of nRMSE scores obtained with the AR model for 1-hour and 6-hour forecast horizons with different studies. Information within parenthesis represents the figure from which values have been visually determined.

Figure 2.11 gives an overall view of the accuracy of forecasts during all sky conditions, but it provides few details concerning the error distribution. To address this shortcoming, let us consider the nRMSE distribution according to the solar elevation angle and the local weather situation represented by the CSI (a low CSI indicates an overcast situation, while a CSI close to 1 represents a sunny situation). Unsurprisingly, Figure 2.12 shows that the forecast error increases as the forecast horizon extends. The graph highlights that the greatest errors are observed when the Sun is at its highest point in the sky, which corresponds to moments with high irradiance levels. The greatest errors are also associated with overcast situations (i.e. low CSI). Both models perform well during sunny situations (i.e. CSI close to 1) and for very low solar elevation angles, for which irradiance levels are very low. We observe that the RF model tends to uniformly reduce forecast errors for all weather situations and elevation angles compared to the AR-based forecasts.

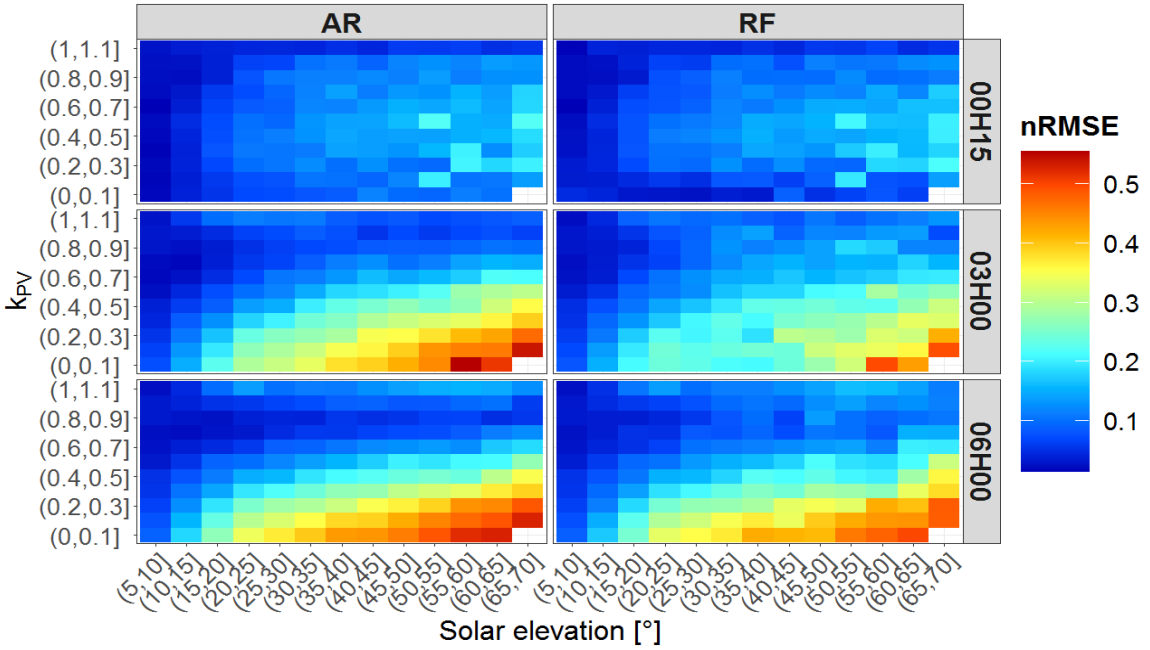



Figure 2.12 – Regime-dependent nRMSE scores obtained with forecasts issued by the AR and RF models as a function of classes of CSI, $k_{PV}$ (defined in Equation 4.16), and solar elevation angles for PV3.

As a complement to our previous forecast quality analysis, we consider a visual implementation of the distribution-oriented framework developed in Section 2.3.4. Figures 2.13 and 2.14 represent the joint and marginal distributions of generation observations for several forecast horizons based on the AR and RF models respectively. We only focus on the analysis of Figure 2.13 as both graphs are rather alike. First, we observe that overall, for the very first time-step, the scatter points are centred on the identity line, but a closer examination of the 2D kernel density in low-production conditions reveals that forecasts slightly drift above the identity line. For higher forecast horizons, the joint distributions are no longer centred on the identity line. Predictions tend to be higher than observations

for low-production levels, but an opposite tendency is observed for high-production rates. This over-estimation of low-irradiance levels may be explained by the fact that forecasts for the early morning (i.e. generated during nighttime) represent mean behaviour learnt during the training phase in the absence of previous production observations. This issue is fixed in Section 5.4 by considering nighttime observations or NWPs model outputs of the irradiance. Histograms represent the marginal distributions of forecasts (on the right) and observations (at the top). The histogram of forecasts for 15-min ahead corresponds well with the observational data distribution (i.e. both histograms exhibit two maximums at low- and high-irradiance rates). However, we observe the appearance of new maximums at mid-irradiance rates for higher forecast horizons, in so much as the marginal distribution of 6-hour ahead forecasts is very different from the production observations histogram.

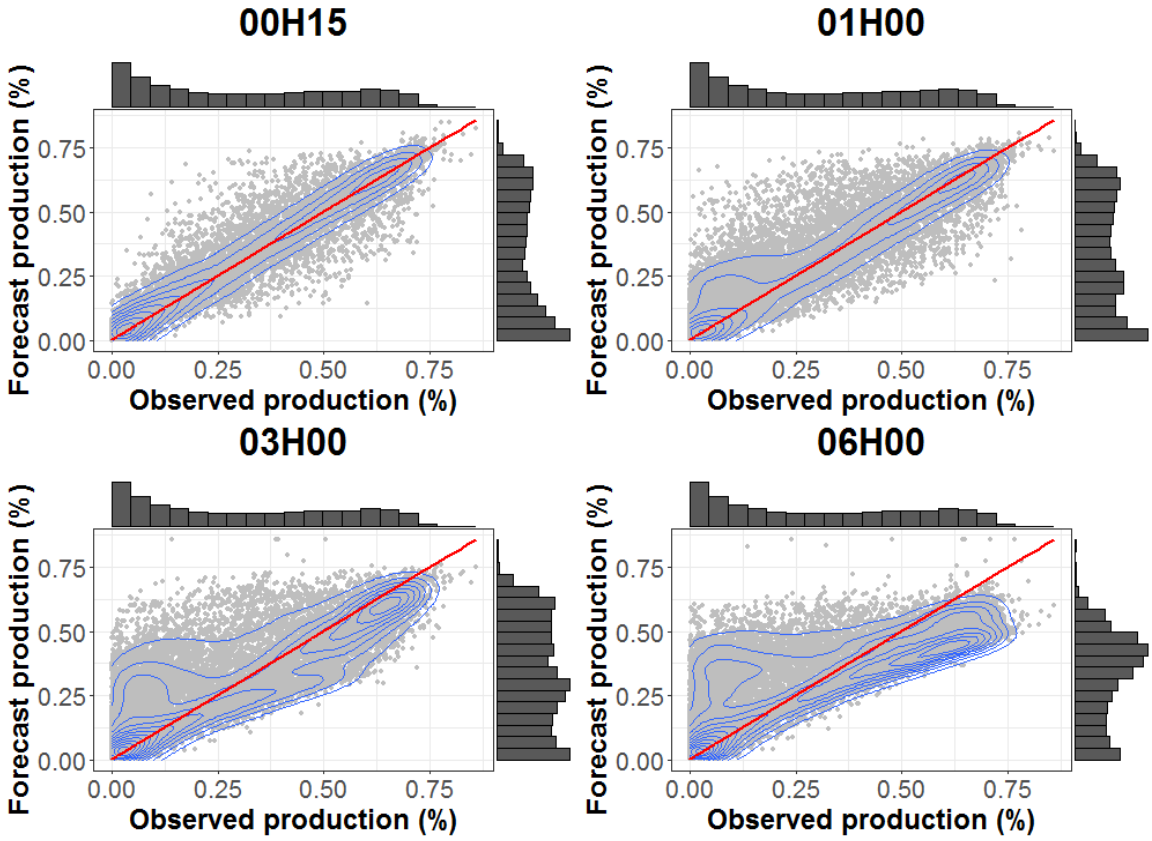



Figure 2.13 – Joint and marginal distributions of 15-min, 1-hour, 3-hour, and 6-hour ahead forecasts and production observations considering the AR model at PV1. The contour lines represent the 2D kernel densities. The red line represents the first bisector. Marginal plots constitute histograms of forecasts and observed production.

## 2.6 Conclusions

This chapter lays the foundation of a modular forecasting framework, which is extended throughout this thesis. The kernel of this architecture is a regression model chosen for its proven performance, low computational cost, low complexity, and interpretability capabili-

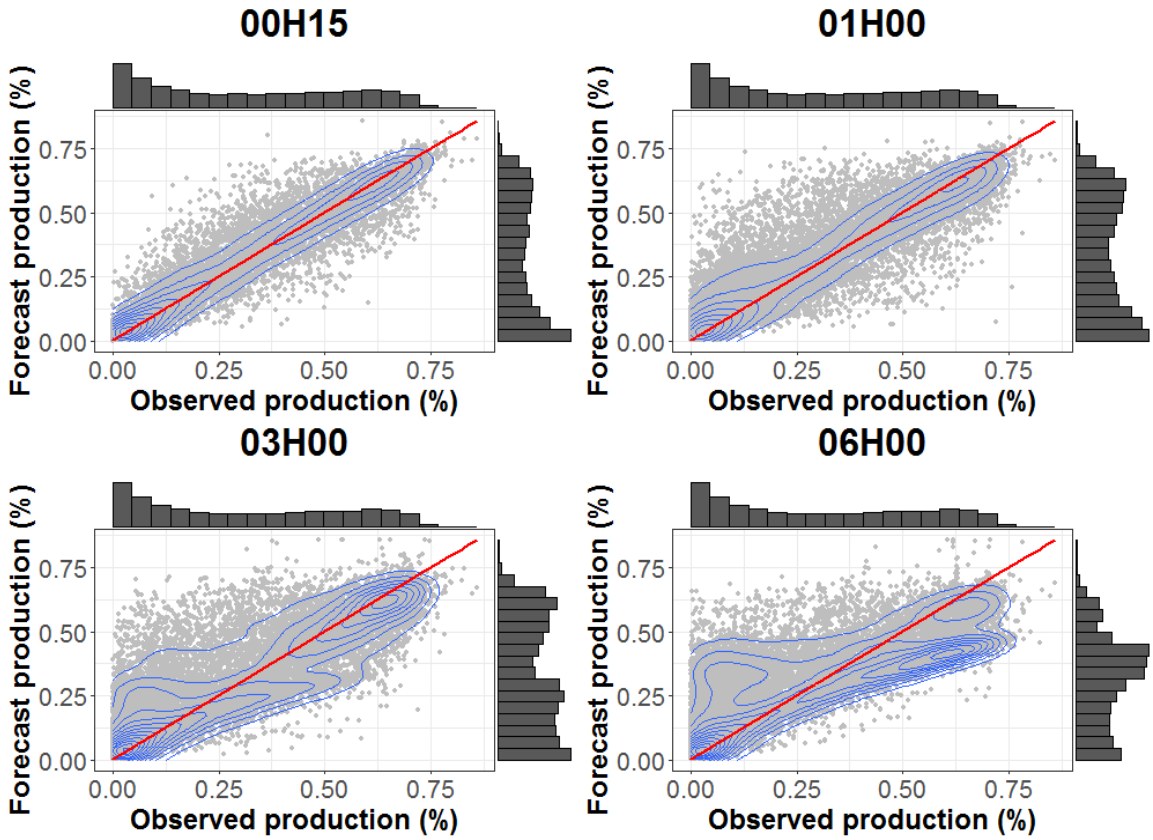



Figure 2.14 – Joint and marginal distributions of 15-min, 1-hour, 3-hour, and 6-hour ahead forecasts and production observations considering the RF model at PV1. The contour lines represent the 2D kernel densities, while the red line is the first bisector of the graph. Marginal plots constitute histograms of forecasts and observed production.

ties.

We adopt a validation framework recommended by eminent solar forecasters to provide in-depth accuracy analysis of forecasts, and to facilitate comparison with other studies. This latter comparison is based on a quantitative and qualitative analysis. Accuracy skill scores provide a good indicator of the global skilfulness of the models and allow cross-work comparisons, while joint and marginal analyses of forecasts compared with observations enable us to assess the error distribution.

To improve forecast accuracy, we also consider a pool constituted by several data sources, namely past PV production observations, satellite-derived observations, and NWPs model outputs. This choice is in line with the observed forecast paradigm shift from temporal-based forecasting to ST-based forecasting. This study goes beyond what can be generally found in the literature inasmuch as infrared sources of information are investigated in Chapter 5.

Lastly, the proposed methodology is implemented to illustrate the forecasting and validation frameworks. These preliminary results act as baseline forecasting performances, which are improved incrementally throughout this work.

## 2.7 Résumé en Français

Pour prévoir la production PV nous utilisons une architecture modulaire que nous développons de manière incrémentale au fur et à mesure de ce manuscrit. Les modules fondamentaux utilisés restent invariants et sont (1) un module propre aux données d'entrée, (2) un modèle de normalisation permettant de s'affranchir de la composante déterministe du signal de production/d'irradiance et de ne garder que la composante stochastique associée aux déplacements de masses atmosphériques, (3) un modèle de régression qui constitue le cœur de cette chaîne de modélisation, et enfin (4) un modèle permettant de dé-normaliser la sortie du modèle de régression et d'obtenir une prévision équivalente à une puissance électrique (Figure 2.2). Tous les modèles sont implémentés et développés via le langage de programmation R.

### Modèles de prévision

Sélectionner le modèle de prévision optimal pour une application spécifique n'est pas une tâche aisée dans la mesure où aucun ne se différencie dans tous les domaines et qu'une profusion importante de modèles est présente dans la littérature. Pour guider notre choix, nous considérons des critères tels que la complexité, l'interprétabilité ou encore l'extensibilité des modèles.

Dans le domaine de la prévision PV, les modèles ARIMA constituent la famille la plus représentée en raison notamment de leur utilisation en tant que modèles de référence mais également en raison de la précision de leurs prévisions, de leur interprétabilité aisée et de la rapidité de leurs calages. Puisque dans les prochains chapitres nous serons amenés à diversifier et augmenter les données d'entrée, il est judicieux de considérer une approche de sélection des variables afin d'éviter des problématiques de surapprentissage et d'améliorer l'interprétabilité des modèles générés. De ce fait, nous implémentons la procédure LASSO avec le modèle auto-régressif (AR).

Puisque le modèle AR est un modèle linéaire, ce dernier n'est pas en mesure de capturer l'ensemble des dynamiques atmosphériques. Un modèle plus avancé est donc nécessaire. Typiquement, dans la littérature dévolue à la prévision de la production PV, deux grandes familles de modèles sont représentées : les réseaux neuronaux profonds (DNN) et les modèles dérivés des arbres de décisions. Ce phénomène peut s'expliquer par la popularité sans cesse croissante des architectures neuronales et leur large gamme d'applications et par le fait que les modèles dérivés des arbres de décisions se retrouvent bien souvent en haut du classement des compétitions de prévision. Dans le cadre de ces travaux, nous avons choisi de travailler avec le modèle RF principalement pour la simplicité de son calage en comparaison à d'autres modèles tels que les GBRT.

### Cadre d'évaluation des performances

Pour évaluer les prévisions générées par les modèles de régression, il est nécessaire de répondre à la question : « qu'est qu'une bonne prévision ? ». Les réponses peuvent être variées selon le contexte d'étude. Ici nous considérons qu'une bonne prévision est une prévision proche de la réalité observée, en d'autres termes, que l'erreur entre les deux variables est faible.

Nous adoptons un cadre de validation des résultats prôné par la littérature. Celle-ci fournie une large gamme de métriques pour quantifier les erreurs de prévision. Ici nous n'en retiendrons que trois, en l'occurrence, lanRMSE, la nMAE, et la nMBE. Pour juger de la performance d'un modèle au regard d'un autre, nous considérons le score de compétence défini à l'Equation 2.9. Enfin, nous considérons la persistance de l'indice ciel clair comme modèle de référence afin de fournir un point de comparaison avec la littérature. En quelques mots, ce modèle suppose que la situation nuageuse à un instant $t$ perdure jusqu'à l'instant $t + h$.

Nous verrons par la suite que, pour certaines configurations de modèles, il est possible d'obtenir des performances prédictives si proches qu'il en devient difficile de savoir si les différences sont réellement significatives. Pour pallier ce problème, nous nous tournons vers le test statistique de Diebold-Mariano qui compare la précision de deux modèles de prévision.

### Données d'entrée

Pour alimenter nos modèles de prévisions, nous considérons trois sources hétérogènes d'information, à savoir l'historique de production des centrales PV, des informations obtenues à partir de satellites, et des prévisions numériques du temps (NWPs).

Tout d'abord, nous avons à disposition l'historique de production de neuf centrales PV réparties dans le sud-ouest de la France, principalement le long du Rhône. Malgré leur localisation dans une région spécifique , elles sont néanmoins soumises à des conditions climatiques diverses : les parcs les plus au Sud subissent un climat méditerranéen alors que la centrale la plus à l'Est est influencée par un climat alpin.

De plus, l'usage de données d'origine satellitaire est également investigué. Cette source de données se décline majoritairement sous la forme d'estimations de l'irradiance au sol (SDSI). Cependant, lorsque les prévisions sont générées pendant la nuit, ce type de données ne fournit aucune information pertinente. C'est pourquoi nous nous tournons vers des cartes d'opacité obtenues à partir des canaux infrarouges des satellites. A notre connaissance, ce type de données est très peu utilisé dans la littérature : nous avons identifié uniquement deux articles les utilisant ([104, 114]). Les images obtenues à partir de satellites offrent l'avantage de couvrir une région étendue autour de la centrale PV et donc renseignent sur les perturbations météorologiques à venir. Le recours à ce type de données fait partie d'un changement de paradigme que l'on observe depuis plusieurs années, et qui consiste à utiliser des méthodes valorisant les dépendances ST entre les données.

Jusqu'à présent, nous avons considéré des données observationnelles, désormais tournons-nous vers des prévisions. Les prévisions météorologiques sont obtenues à partir de modèles numériques de l'atmosphère représentant avec plus ou moins de simplifications les différents processus physiques à l'œuvre. Dans cette étude, nous considérons deux types d'information : des prévisions numériques au niveau du site d'intérêt (i.e. la moyenne des variables obtenues à partir des quatre points de grille les plus proches) ou des prévisions sous forme de carte 2D centrées au niveau des centrales. Ce premier type d'information est constitué de variables sensibles telles que l'irradiance, la température ou encore la couverture nuageuse totale. Le second type d'information quant à lui représente une variable synoptique, en l'occurrence le champ géopotentiel. Cette variable fournie de précieuses informations concernant le type de situations météorologiques à l'œuvre dans la région concernée (e.g. temps ensoleillé, orageux) ainsi que des renseignements sur la direction des déplacements des masses nuageuses.

### Résultats préliminaires

La dernière partie de ce chapitre présente quelques résultats préliminaires concernant des prévisions obtenues uniquement à partir de l'historique de production. Dans ce cas de figure particulier, le modèle RF surpasse le modèle AR pour l'ensemble des trois métriques considérées. Néanmoins, les performances obtenues en considérant le modèle AR sont en accord avec ce qui est observé dans la littérature. Cette partie est également l'occasion d'introduire les outils graphiques utilisés pour analyser les performances des modèles.

# Chapter 3

# Physics-based Modelling

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.
>
> Pierre Simon Laplace, *A Philosophical Essay on Probabilities* (1814)

**Contents**

## 3.1 Introduction

Alternative Current (AC) power generated by a Photovoltaic (PV) power plant is the result of a complex conversion process involving several components affected by specific environmental and technological factors, the best example being the geometry-dependency (i.e. orientation and inclination of modules) of PV production.

Several approaches are conceivable to derive production forecasts. The PV generation forecasting field can be split into three main categories: (1) physical models, (2) statistical models, and (3) hybrid models (i.e. a combination of both previous categories). The first approach explicitly defines the conversion processes involved via physical laws (as such it can be considered as a white box), while the second family uses statistical models to infer these laws (i.e. grey- or black-box modelling depending on the model's complexity). In truth, the physical modelling class is not a set of forecasting techniques in itself inasmuch as the forecasting effort is supported by Numerical Weather Predictions (NWPs) models. This class of model is not widespread in the PV forecasting literature in comparison with the statistical modelling family [47]. To account for this under-representation, [115] points out the lack of data regarding plants' design parameters.

The main objective of this section lies in investigating the potential interest of linking physics- and statistics-based modelling. A rich Photovoltaic Production Forecasting (PVPF)-related literature has been developed to account for physical effects intervening in the irradiance-to-electricity conversion process [115], while usually statistical models neglect this aspect for the benefit of better modelling of weather phenomena. This objective can be viewed from a different perspective, namely, a way to integrate explicitly physics-based knowledge within statistical models with the aim of reducing the inferring effort. Such an ambition raises the following questions:

**Research Gap - *Global Horizontal Irradiance conversion***

In the literature, most studies based on Machine Learning (ML) or Deep Learning (DL) tools consider the irradiance information in the form of Global Horizontal Irradiance (GHI). Such an approach postulates that the transposition into the Plane-of-Array (POA) [a] is implicitly performed by the forecasting algorithm. Is it possible to reduce the modelling efforts upon forecasting models and improve forecasting performances by considering Global Tilt Irradiance (GTI), or even the electrical power derived from NWPs? If so, what are the critical modelling steps?

*a.* Plane of array irradiance quantifies the incident irradiance on a given solar array.

## 3.2 Methodology

To fulfil the aim of this chapter, it is necessary to develop a modelling chain able to convert a set of inputs composed of weather variables and plant-specific parameters into electric power. This physics-based chain converts the GHI data obtained from clear-sky, satellite-based, or NWPs models, and even pyranometers observations, into electric power. The resulting preprocessed irradiance is then injected into a forecasting model. This approach offers the possibility to include explicitly physics-based knowledge in statistical models and, therefore, leads us to develop a hybrid modelling strategy.

For each modelling step, a wide range of models is found in the literature. To limit the choice of possible options, the design strategy of this conversion chain has to obey two main criteria: simplicity and accuracy. The featured models must consider easily retrievable inputs to enable large-scale use and little computing efforts.

A literature review identified the prevailing modelling steps within the scope of PVPF. These processes are summarised in Table 3.1. Neglected processes may be implicitly taken into consideration in the statistical forecasting model by providing relevant inputs (e.g. ageing can be assessed if a time-based feature is provided).

| **Physical modelling** | **Neglected processes** |
|---|---|
| GHI decomposition | Electricity conversion |
| GHI projection | Angular losses (dust and soiling) |
| Shading effects (inter-row and far shading losses) | Angular losses (spectral response) |
| Angular losses (reflection) | Ageing |
| Irradiance-to-power modelling | Shading effects (near shading losses) |
| Cells temperature | |

Table 3.1 – Explicitly modelled or rejected processes intervening in the conversion of irradiance into electrical power.

To give the reader more insight into the physics-based modelling chain, Figure 3.1 represents the general architecture of the conversion process retained. The steps impacting irradiance before it reaches the PV cells are detailed in Section 3.3, while Section 3.4 describes the conversion of irradiance into electrical power.

To evaluate the impact of the physics-based modelling on forecasting performances, two strategies are adopted. First, performances are assessed within the clear-sky normalisation framework (process defined in Section 4.4). In this sense, this work is an extension of results provided in Section 4.4.2.2.3. Second, it is common practice in the PVPF field to resort to non-normalised data (i.e. data which are not normalised by clear-sky-based features) when using ML tools. Thus, it seems interesting to compare the forecasting performances of models fed respectively with raw irradiance and preprocessed irradiance.

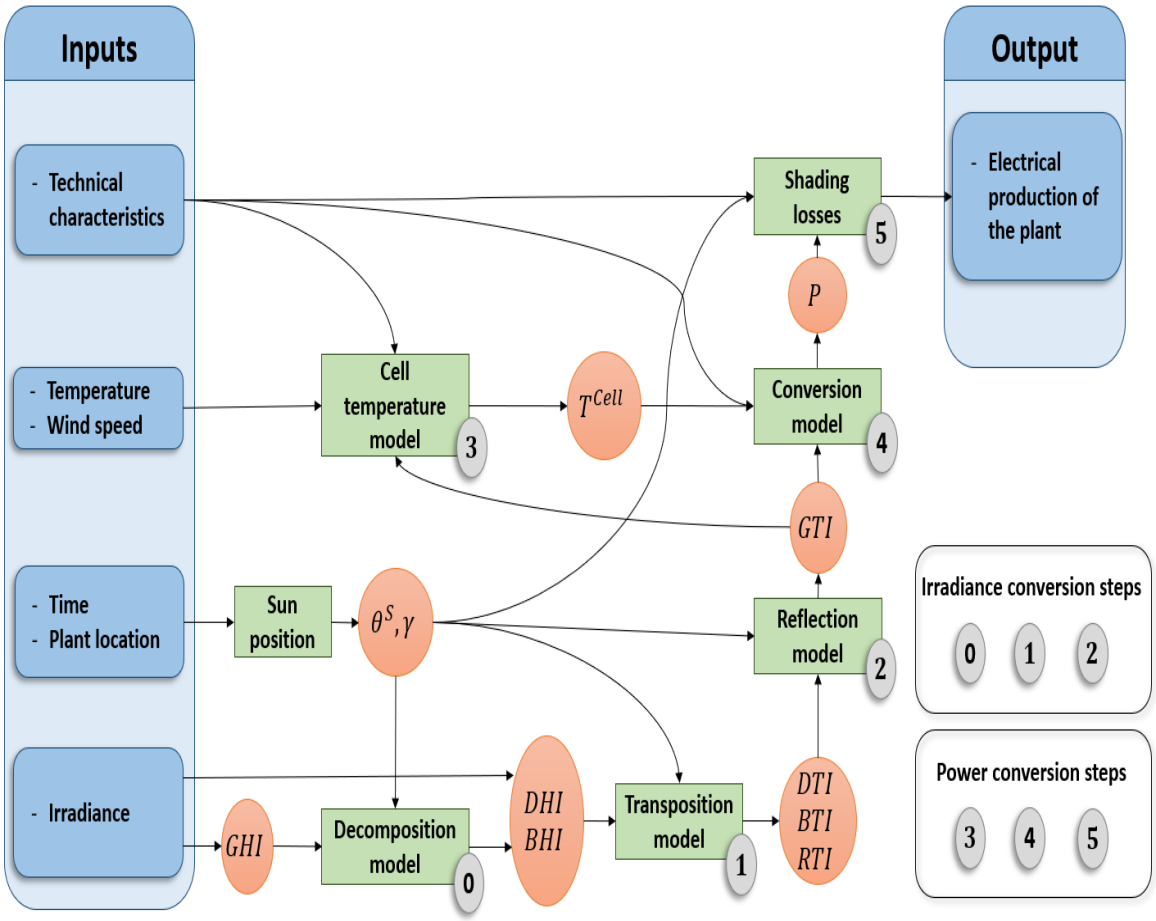



Figure 3.1 – Physical PV production conversion chain composed of a decomposition model (3.3.1.1), a transposition model [a] (3.3.1.2), a reflection model (3.3.3), a cell temperature model (3.4.2), a power conversion model (3.4.4), and the modelling of shading losses (3.3.2) (graph based on [115]).

*a.* Step (0) may be optional if the components of irradiance are provided.

## 3.3 Effective irradiance reaching photovoltaic cells

When dealing with short-term PV generation forecasting, apart from past production observations, irradiance data are the main impacting factor. This feature is usually provided in the form of GHI by numerical models, which take into account the influence of the atmosphere and its components. It is then necessary to transpose the horizontal irradiance to the tilted plane of the PV modules to obtain the GTI. After crossing the atmosphere, the solar irradiance has to cross the panel cover before reaching the PV cells. At this point, the light beam can be altered by angular reflection and the presence of dust on the glass cover.

Figure 3.2 illustrates the geometrical angles used in mathematical formulations throughout this section.

$\theta^S(t)$ Solar zenith angle formed by the direction of the Sun and the local vertical (°, unless otherwise specified),

$\gamma^S(t)$ Sun's elevation angle (i.e. $\theta^S(t) + \gamma^S(t) = 90°$) (°),

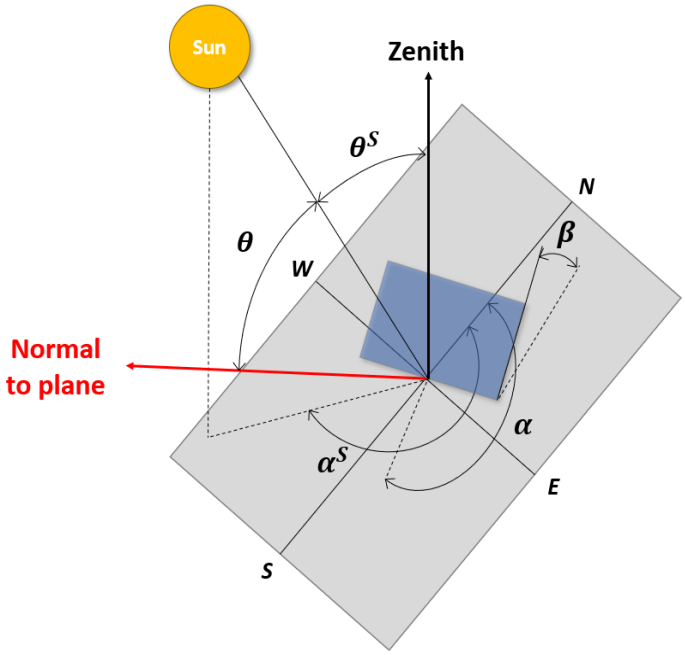



Figure 3.2 – Angles describing the position of the Sun and the panel positioning. N, E, S, W denote the north, east, south and west. Inspired from [116].

- $\alpha^S(t)$ Solar azimuth angle (°),
- $\theta(t)$ Incident angle: angle comprised between the normal to the plane and the solar rays (°),
- $\beta$ Inclination angle of the panel (°),
- $\alpha$ Azimuth angle of the panel, i.e. angle between the projection of the normal to the plane and the north direction (°).

### 3.3.1 Irradiance on the plane-of-array

#### 3.3.1.1 Decomposition of the GHI

The incoming solar radiation reaching a horizontal plane on the ground is composed of two main components, namely the direct solar radiation (i.e. Beam Horizontal Irradiance (BHI)) and the diffuse solar radiation (i.e. Diffuse Horizontal Irradiance (DHI)) (Equation 3.1). These parameters are crucial to derive solar radiation on a tilt plane or to determine shading losses.

$$GHI(t) = BHI(t) + DHI(t) \tag{3.1}$$

In the present study, the McClear model outputs [117] and the European Centre for Medium-Range Weather Forecasts (ECMWF) numerical model [1] [118] provide respectively

1. The NWPs model provides the SSRD (surface solar radiation downwards) and FDIR (total sky direct solar radiation at surface) parameters, which correspond to the GHI and BHI quantities. A simple computation leads to the diffuse solar radiation.

estimations and predictions of BHI and DHI components. On the contrary, the Satellite Derived Surface Irradiance (SDSI) maps database only provides GHI quantity. Consequently, in this latter case, it is necessary to resort to a separation (or decomposition) model to estimate the beam and diffuse components from the GHI [2].

Most of the decomposition models are empirically derived from irradiance measurements [119] and aim at estimating the diffuse fraction, $k_D$, (i.e. the ratio of the diffuse to the GHI, Equation 3.2) as a function of the clearness index, $k_t$ (i.e. the ratio of the GHI to the horizontal extraterrestrial irradiance) and other predictors (e.g. air mass, dew-point temperature, relative humidity, zenith angle) [115]. Gueymard in [119] proposes a comprehensive evaluation and validation study of around 140 separation models for five climatic regions, which provides two interesting conclusions:

1. The *ENGERER2* model [120] can be considered as a 'quasi-universal' 1-min separation model, wherever and whenever low-albedo conditions prevail,
2. The most recent models do not generally offer improved accuracy in comparison with the first proposed separation models except for the local area that they are specifically designed for.

$$k_D(t) = \frac{DHI(t)}{GHI(t)} \tag{3.2}$$

This *ENGERER2* model has been specially developed to account for Cloud Enhancement (CE) situations. The CE phenomenon occurs on partly cloudy days when the irradiance temporarily exceeds the expected clear sky irradiance value. This situation is assumed to result from reflections from cloud edges and strong forward Mie scattering inside the cloud [121], which increase the diffuse part of irradiance. Typically, CE only appears in high-resolution measurement data as it lasts from seconds up to a minute [121]. Thus, a model as refined as the *ENGERER2* model is not well suited to work with 15-min resolution data. Therefore, we turned to the Boland–Ridley–Lauret (BRL) decomposition model developed in [122] as it is considered as one of the best separation models [115, 123] and does not require measured or forecast predictors, but only variables, which can be easily computed with solar geometry algorithms (Equation 3.3 and Equation 3.4). It is worth mentioning that initially the model was developed with hourly data, but [124] proposes verifying the usefulness of the model for minute data.

$$k_D(t) = \frac{1}{1 + e^{-5.38+6.63k(t)+0.006AST(t)-0.007(90-\theta^S(t))+1.75K(t)+1.31\psi(t)}} \tag{3.3}$$

$\psi(t)$ is a measure of persistence of global radiation level and $K(t)$ is the daily clearness index:

2. In this work, the decomposition model is used to make up for a lack of data (namely the diffuse and direct irradiance derived from satellite-based devices). In an operational context it can also prove to be relevant for cost reduction by avoiding the need to buy additional forecast products (i.e. diffuse irradiance), but this may be to the detriment of accuracy.

$$\psi(t) = \begin{cases} \frac{k(t-1)+k(t+1)}{2} \text{ between sunrise and sunset,} \\ k(t+1) \text{ at sunrise,} \\ k(t-1) \text{ at sunset.} \end{cases} \qquad K(t) = \frac{\sum_{j=1}^{N'_d} GHI(j)}{\sum_{j=1}^{N'_d} I^{0h}(j)} \tag{3.4}$$

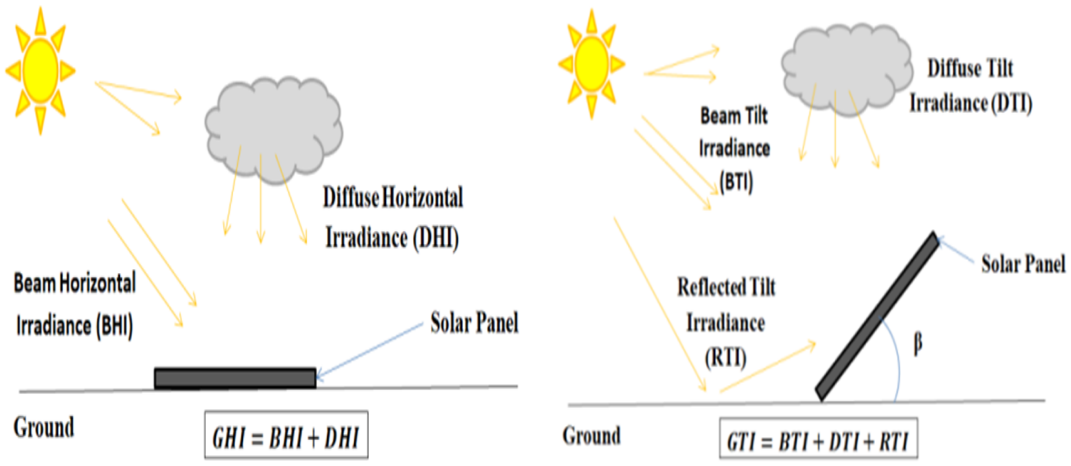
Where $I^{0h}(t)$ is the horizontal extraterrestrial irradiance (Equation 3.5), and $I^{0n}(t)$ is the normal extraterrestrial incidence irradiance (Equation 3.9).

$$I^{0h}(t) = I^{0n}(t) cos(\theta^S(t)) \tag{3.5}$$

| | |
|---|---|
| $k(t)$ | Clearness index ($\emptyset$), |
| $AST(t)$ | Apparent solar time (hour), |
| $\theta^S(t)$ | Solar zenith angle (°), |
| $K(t)$ | Daily clearness index ($\emptyset$), |
| $N'_d$ | Number of observations during the day ($\emptyset$), |
| $\psi(t)$ | Clearness index persistence ($\emptyset$). |

#### 3.3.1.2 Projection of the GHI

Now that the DHI and BHI are known, it is possible to project incoming irradiance (i.e. the GHI) on the POA to derive the GTI (Figure 3.3).




(a) Ground Horizontal Irradiance (GHI). (b) Ground Tilt Irradiance (GTI).

Figure 3.3 – Illustration of the GHI and the GTI.

The first transposition (or projection) models appeared in the 1960s. This family of models estimates the irradiance on a tilted surface with arbitrary orientation from the horizontal irradiance data. Generally, in the literature, we found two distinct groups of models: (1) physics-based approaches, which model the irradiance in the sky, and (2) machine learning-based methods (e.g. [125]), which learn the transposition relationship between the GHI and the GTI. In this document we focus on physics-based approaches. The global tilted

irradiance is the sum of the beam (or direct irradiance, but the term *beam* is preferred to avoid any confusion when acronyms are used), the diffuse, and the ground-reflected components over the plane of array. The projection of the horizontal components is modelled via transposition factors (Equation 3.6).

$$GTI(t) = BTI(t) + DTI(t) + RTI(t), \text{ with: } \begin{cases} BTI(t) = BHI(t) \cdot R^b(t) \\ DTI(t) = DHI(t) \cdot R^d(t) \\ RTI(t) = \rho \cdot GHI(t) \cdot R^g \end{cases} \tag{3.6}$$

**3.3.1.2.1 Diffuse component** The main difference between the various transposition models available in the literature lies in the way diffuse irradiance is computed [115, 126]. In [127], the authors classify the methods for calculating the diffuse tilted irradiance into three categories: (1) methods assuming an isotropic sky, (2) models based on an an-isotropic sky by considering circumsolar radiation (i.e. radiation from the bright region surrounding the solar disc), and (3) models including a horizon brightening component. Such models represent idealised cases, where the foreground is assumed to be un-shaded and infinite. A wide variety of projection models are available in the literature. In this regard, [126] provides a fairly comprehensive review and benchmarks twenty-six models: the authors identify the Perez family of models as the overall best performer. This transposition model splits the sky hemisphere into three areas: the circumsolar disc, the horizon band and the isotropic background. We choose to consider the model [3] presented in [128] and defined hereafter. To characterise the weather situation, the sky's clearness, $\epsilon(t)$, and the sky's brightness, $\Delta(t)$ are used:

$$\epsilon(t) = \frac{\frac{DHI(t)+BNI(t)}{DHI(t)} + 1.041 \left(\frac{\pi}{180}\theta^S(t)\right)^3}{1 + 1.041 \left(\frac{\pi}{180}\theta^S(t)\right)^3} \qquad \Delta(t) = \frac{DHI(t) \cdot AM(t)}{I^{0n}(t)} \tag{3.7}$$

$AM(t)$ is the relative optical air mass [4] [129] (Equation 3.8) and $I^{0n}(t)$ is the normal extraterrestrial incidence irradiance [130] (Equation 3.9) where $I_0$ is the solar constant (here $I_0 = 1361\ W/m^2$ [116]), and $n_t$ is the day of the year.

$$AM(t) = \frac{1}{cos(\theta^S(t)) + 0.50572\left((90 - \theta^S(t)) + 6.07995\right)^{-1.6364}} \tag{3.8}$$

$$I^{0n}(t) = I_0 \left(1 + 0.033 cos\left(\frac{360 n_t}{365}\right)\right) \tag{3.9}$$

3. The Perez model is derived from 13 sites located in Switzerland, France and the USA (climatic environments of experimental datasets are provided in Table 2 from [128]).

4. This is the ratio between the length of the optical path through the atmosphere of solar radiation and the path length at the zenith. This quantity is dependent on the Sun's position and allows us to characterise the solar spectrum.

The diffuse transposition factor,$R^d$, is determined with Equation 3.10. $F_{i,j}$ parameters depend on the sky's clearness, $\epsilon$, (see Table 1 and Table 6 from [128] to obtain the value of the function depending on the discrete sky's clearness categories).

$$R^d(t) = \underbrace{\left(1 - F_t^1(\epsilon(t), \Delta(t), \theta^S(t))\right) \cdot F_{vf}}_{\text{Isotropic component}} + \underbrace{F_t^1(\epsilon(t), \Delta(t), \theta^S(t)) \cdot \left(\frac{a}{b}\right)}_{\text{Circumsolar component}} + \underbrace{F_t^2(\epsilon(t), \Delta(t), \theta^S(t)) \cdot sin(\beta)}_{\text{Horizon brightening component}}$$

$$\begin{cases} F_{vf} = \dfrac{1 + cos(\beta)}{2} \cdot F_t^1(\epsilon(t), \Delta(t), \theta^S(t)) \\ \qquad = F_{11}(\epsilon(t)) + F_{12}(\epsilon(t)) \cdot \Delta(t) + F_{13}(\epsilon(t)) \cdot \theta^S(t) \\ F_t^2(\epsilon(t), \Delta(t), \theta^S(t)) = F_{21}(\epsilon(t)) + F_{22}(\epsilon(t)) \cdot \Delta(t) + F_{23}(\epsilon(t)) \cdot \theta^S(t) \\ a = max(0, cos(\theta(t))) \\ b = max(cos(85°), cos(\theta^S(t))) \end{cases} \tag{3.10}$$

It is worth mentioning that this model considers only one row of PV modules. Thus, when dealing with a series of adjacent parallel rows of panels, the soft shading (i.e. the reduction of diffuse light due to adjacent lines) is not taken into account. In the same manner, the foreground is usually assumed to be infinite, but in practice, ground reflection on modules outside the front row may be overestimated.

**3.3.1.2.2 Direct and reflected components** The projection of the beam component, Beam Tilt Irradiance (BTI), is based on geometry [131], while the ground-reflected component, Reflected Tilt Irradiance (RTI), is obtained by considering isotropic irradiance and the ground albedo, $\rho$ [132] (Equation 3.11).

$$R^b(t) = max\left(0, \frac{cos(\theta(t))}{cos(\theta^S(t))}\right) \qquad R^g = \frac{1 - cos(\beta)}{2} \tag{3.11}$$

With the angle of incidence, $\theta$, defined as [130]:

$$cos(\theta(t)) = cos(\theta^S(t)) \cdot cos(\beta) + sin(\theta^S(t)) \cdot sin(\beta) \cdot cos(\alpha^S(t) - \alpha) \tag{3.12}$$

| | |
|---|---|
| $\theta(t)$ | Angle of incidence: angle between the beam radiation on a surface and the normal to that surface, |
| $\theta^S(t)$ | Solar zenith angle, |
| $\alpha^S(t)$ | Solar azimuth angle, |
| $\beta, \alpha$ | Inclination and azimuth angles of the panel, |
| $\rho$ | Ground albedo ($\rho = 0.2$). |

At this point, the irradiance reaching the PV module's cover has been quantified. Let us focus on the losses generated during the crossing of the glass cover of the module.

### 3.3.2 Shading effects

#### 3.3.2.1 Inter-row shading

The partial shading of PV devices can result from weather phenomena (i.e. displacement of clouds, snow, dust) or from nearby structures that alter sunlight (i.e. neighbouring solar panels, buildings or trees). In this section, we focus on this last category and more specifically on adjacent panel rows.

Ground-mounted PV plants with multiple parallel structure rows may induce energy losses owing to adjacent rows creating shade over each other. Inter-row shading depends on the inter-row distances and is observed for low solar elevation angles (i.e. mainly during winter time, sunrise and sunset). In this regard, row spacing is typically determined during the planning phase of the solar plant to avoid shading at noon at winter solstice [127], but [133] put forward that it is not rare to find plants with low spacing between rows, which leads to significant row-to-row shading effects. Such configurations may originate from high land costs combined with low module prices.

Inter-row shading results in unevenly distributed irradiance on the plane of PV modules: (1) the un-shaded part benefits from direct, diffuse and ground-reflected solar irradiance, while (2) a part of this irradiance is blocked for shaded regions. The shading from adjacent mounting structures is purely geometrical, while the associated mismatch losses (notion defined in Focus 3.1) depend on the network architecture (e.g. presence of bypass diodes, implementation level of Maximal Power Point Tracking (MPPT)) and the interconnection schemes (e.g. series-parallel, total-cross-tied) [127]. Usually, the proportion of electric shading losses is higher than the decrease in irradiance because a single shaded cell can limit the current of all the series-connected cells [134]. Partial shading can also prevent the system from operating at the Maximum Power Point (MPP): due to non-uniform irradiation levels, multiple MPP are available on the P-V characteristics, which challenges conventional MPPT to track the global maximum [135]. As a result, to model inter-row shading losses it is necessary to have complete knowledge about the connection and arrangement of the modules. Besides, the PV fields are rarely perfectly flat (Figure 4.2). Such topographies lead to some bias because the following equations accounting for shading effects are designed for horizontal planes. For the present study, we consider that the required modelling complexity to account for such effects is not justified by the added modelling accuracy.

In [133], the authors propose a fairly simple formulation of the shaded fraction of the module area as a function of the site configuration and the Sun's position. This approach assumes that the beam irradiance is the only component affected by the light obstruction from adjacent row panels. In fact, the diffuse and reflected components are also affected: e.g. the lower portion of the sky is obscured by adjacent rows which modify the isotropic diffuse transposition factor presented in Equation 3.10. To account for this phenomenon, it is possible to replace the term $F_{vf}$, which represents the view-factor of the first row, by the view-factor proposed in equation (1) in [138] (this expression can be applied to rows in a PV

**Focus** 3.1 – **Mismatch effects**

A PV module is usually composed of a set of 60 PV cells connected in a series. This configuration increases the entire voltage output of the module, while the same current value flows through all the cells. The series connection forces all the cells in a string (i.e. a set of components connected in a series) to work at the same current, which leads to mismatch losses if one cell is operating at a different point. In simple terms, the mismatch loss can be defined as the difference between the expected and actual output power from a PV module. This operating point is determined by the cell's physical properties resulting from the manufacturing process (cells are binned during module fabrication) and from external conditions (i.e. irradiance, temperature). Thus, heterogeneous irradiance distribution reaching the PV module (due to partial shading, and even dust deposition) alters the conversion efficiency of a PV module - and to a larger extent - of a string of modules. Thus, if one cell in the string receives less irradiance than the others, the maximum current of the module is that of the shaded cell (i.e. the module can be considered as shaded). Such a situation can be challenging: indeed, unevenly distributed irradiance affects the pattern of power-voltage characteristic curves of each module, which results in multiple maximum points, increasing the challenge for the MPPT system to find the global power peak point [136]. Besides, mismatch losses can also irreversibly deteriorate and shorten the service life of PV cells due to the rise in temperature. Hot-spot heating appears when the reduced short-circuit current of affected cells becomes lower than the operating current of the module. When such a condition occurs, the affected cell or group of cells is forced into reverse bias, acting as an internal load and dissipating generated power produced by the other *good* cells in the form of heat [137]. To prevent energy losses and cells damage, bypass diodes are connected in reverse parallel with the PV cells to provide an alternative path for the current to flow. Owing to costs concerns, a standard module is usually composed of three sub-strings protected by three bypass diodes [127]. The module is then divided into three groups of 20 cells along its short edge [134].

panel composed of $N$ modules placed one above the other [139]). Depending on the position of the module in the panel, the amount of isotropic diffuse irradiance varies: lower modules are more impacted than higher ones. Thus, such an approach makes the computational chain more complex by requiring the computation of the view-factor for each module. A more advanced model has been presented recently in [127] which performs better than approaches considering only direct shading. This approach takes into account the impact of shading over the diffuse components (e.g. the hard-shading effect over the circumsolar diffuse irradiance) and the reflected components. Yet, including only the beam shading improves the overall modelling accuracy of the irradiance-to-power conversion chain [115]. In the scope of this thesis, we focus only on the beam irradiance shading (Equation 3.13).

$$f_{shading}(t) = \begin{cases} \left[\frac{\left|\frac{L \cdot sin(\beta)}{tan(\gamma^S(t))}\right| - \left|\frac{s}{cos(\alpha^S(t)-\alpha)}\right|}{\left|\frac{L \cdot sin(\beta)}{tan(\gamma^S(t))}\right| + \left|\frac{L \cdot cos(\beta)}{cos(\alpha^S(t)-\alpha)}\right|}\right] & \text{if } |\gamma^S(t) - \beta| < \frac{\pi}{2} \\ 0 \text{ if } |\gamma^S(t) - \beta| \geq \frac{\pi}{2} \text{ or } \gamma^S(t) < 0 \end{cases} \tag{3.13}$$

$\gamma^S(t)$ The Sun elevation angle ($\theta^S(t) + \gamma^S(t) = \frac{\pi}{2}$) (rad),
$L$ Panel length (m),
$s$ Module row interspacing distance (m).

#### 3.3.2.2 Near-shading and far-shading losses

Depending on the location of the plant, the surrounding landscape can be composed of obstructions that cast shade on the PV receivers. Near-shading results from surrounding objects, such as tress or buildings, while far-shading refers to projected shades from distant mountains or hills.

The plants under study are mainly located in rural areas. From Google Earth-based observations, we observe that the PV site surroundings are quite clear, although some are near trees, bearing in mind that even high-voltage lines can shade some outlying panels, but the phenomenon is too limited to justify its modelling.

Most of the PV plants (8 out of 9) are located in the Rhone valley which is surrounded by the Alps to the east and by the Massif Central to the west. As a result, relief heterogeneity is a distinctive feature of this area. To account for distant hills or mountains, the horizon profile for each plant's location is obtained from [140]. The horizon profile is a 360°elevation map representing the profile of the surrounding area computed from elevation measurements (Figure 3.4). If the Sun is below the horizon line, the direct component of irradiance is blocked.

#### 3.3.2.3 Dust and soiling

Soiling on the front glass of the module cover generates additional optical losses due to absorption, scattering and reflection of the incoming sunlight. Soiling has a negative impact on the economic profitability of the PV plants, not only because it reduces the yield, but also because it generates additional cleaning costs [141]. To date, no passive anti-soiling technology completely removes the need for cleaning, which can be performed manually, semi-automatically (e.g. truck-mounted) or fully automatically [141].

Most of the models proposed in the literature rely on experimental data [142], which makes them very environment-dependent. In this area, [143] highlighted that the dust deposition phenomenon has mainly been studied in Middle East region, but few studies have been carried out in Central Europe.

Airborne dust concentration and rain frequency are the main factors affecting soiling. Rainfall seems to have a limited effect on small dust particles (2-10 $\mu$m) but is more efficient

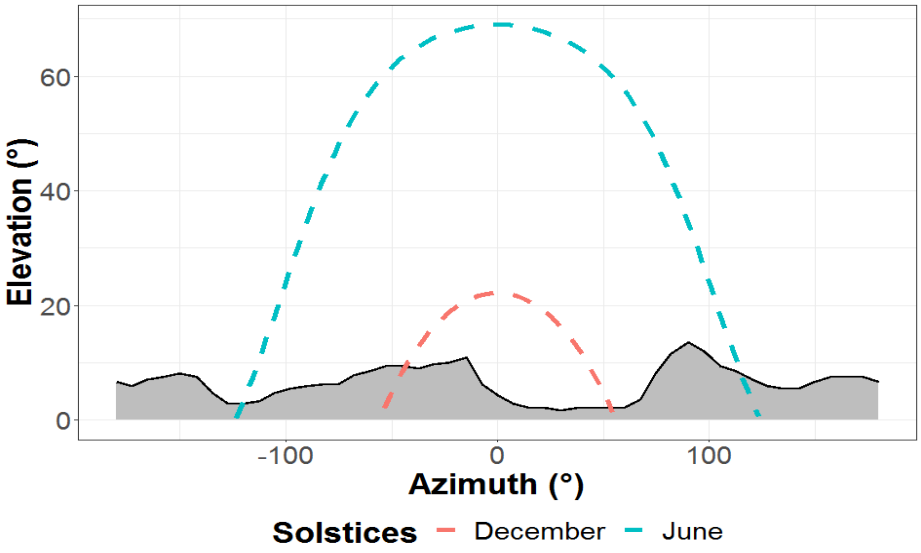



Figure 3.4 – Horizon line at PV10. Blue and red dashed lines represent the Sun's elevation angles at the June and December solstices.

to wash away larger dust particles (e.g. pollen) [143]. In [144], the authors showed that in the south of Spain during summertime, soiling losses are greater owing to less frequent rainfall. While rain is an effective way to clean off dust, wet surfaces resulting from relative humidity strongly enhance dust adhesion [141]. In this regard, several field studies have shown that dust settlement rates decrease as the tilt angle of modules increases.

The dataset under study is composed of PV plants mainly located along the Rhône River. Their surroundings are characterised as a rural background composed of fields, roads, copses and stony pathways. The Compagnie Nationale du Rhône (CNR) does not perform any cleaning activities, considering that rainfall is sufficient. We tried to confirm, a posteriori, this hypothesis by adopting an approach similar to the one found in [145]. A pseudo-performance index is computed as the ratio of the sum of the soiled device power measurements over a day, and the sum of the simulated power over the day. The theoretical power is obtained by considering the model developed in this section with SDSI at the site position. Nevertheless, it was not possible to identify performance degradation over time or performance improvement after rainy events. This is assumed to result from various measurement and modelling errors within the forecasting chain, which are higher than the performance decrease associated with soiling. Therefore, we assume soiling losses to be negligible.

Snow deposition in winter is observed mainly on the easternmost power plant. Days associated with snowfall are rejected (Section 2.4.1).

### 3.3.3 Optical effects

#### 3.3.3.1 Angular losses

A PV module is composed of several layers (i.e. protective cover(s), solar cells), which induce reflection and absorption of solar radiation [146]. Such losses can hardly be neglected as they can reach between 2% and 3% [127]. In [147], the authors emphasised that cell technology has a second order influence over the optical losses, which originate mainly from the reflection of the incident light at the air-glass interface. Although light reflection is wavelength-dependent, the literature showed that this dependence can be neglected [148].

The angular response varies according to the components of the solar radiation.

1. Direct radiation can be characterised by the incidence angle $\theta(t)$. As the incidence angle increases, the amount of reflected light increases to such an extent that significant effects occur at incidence angles higher than 65° [149]. In other words, reflection losses of direct radiation are preponderant when the Sun is low in the atmosphere for PV panels with typical inclination angles.
2. Diffuse and ground-reflected irradiance can be assumed to be isotropic (i.e. radiation intensity is the same and independent from direction), and as such it is necessary to integrate the contribution of each solid angle unit on the PV module. Moreover, the proportion of each component varies according to the Solar Zenith Angle (SZA), $\theta^S(t)$ (e.g. as the SZA increases, the contribution of direct radiation decreases, but the diffuse radiation part increases), and also according to the cloud coverage (in clear-sky conditions, direct radiation is the main contributor to the GTI, while in cloudy situations, the diffuse part of the solar radiation gains more weight).

Thus, to consider the various behaviours, it is necessary to model angular losses with specific formulations for each component.

To account for the optical losses, two types of model can be considered [115]: (1) theoretical models derived from optical laws (e.g. Snell and Fresnel equations), or (2) empirical models. The first class of models aims at accounting for the reflectance and/or the absorption effects occurring at the cover layer of the PV module. With the ever-increasing number of module configurations/technologies, the second family of models is appealing because it provides generic formulations with specific fitted parameters depending on the module's features [147, 150]. Here, we choose to use the theoretical model presented in [151]. This model can be viewed as an extension of the air-glass reflection and absorption model proposed in [149] (for more in-depth mathematical developments, the reader may refer to [152]): the authors included the possibility of considering an Anti-Reflective Coating (ARC) layer and highlighted that the absorption part of the equation can be ignored. This model has been selected because our portfolio of PV plants is composed of 5 plants with ARC and 4 without. To minimise reflection losses, multiple layers of ARC can be applied in combination with surface texturing of solar cells (silicon material possesses a high refractive

index [153]). Here, due to the lack of technical information and for the sake of simplicity, we neglected the influence of texturing and we assume that the ARC possesses only two layers (Figure 3.5).

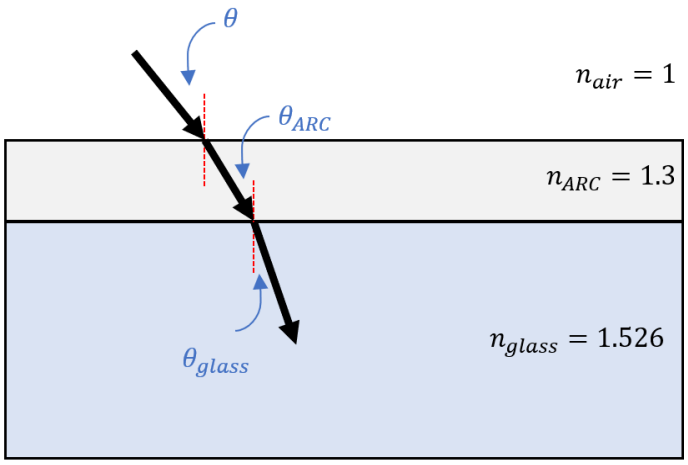



Figure 3.5 – Diagram of the ARC.

To compute the direct radiation optical losses, first, the transmittance through the ARC, $\tau_{ARC}$, is computed with Fresnel's equation:

$$\tau_{ARC}(t) = 1 - \frac{1}{2}\left(\frac{sin^2(\theta_{ARC}(t) - \theta(t))}{sin^2(\theta_{ARC}(t) + \theta(t))} + \frac{tan^2(\theta_{ARC}(t) - \theta(t))}{tan^2(\theta_{ARC}(t) + \theta(t))}\right) \tag{3.14}$$

Where, the angle of refraction into the ARC, $\theta_{ARC}(t)$, is determined from Snell's law:

$$\theta_{ARC}(t) = arcsin\left(\frac{n_{air}}{n_{ARC}} sin(\theta(t))\right) \tag{3.15}$$

Then, the transmittance through the glass, $\tau_{glass}$, is calculated similarly:

$$\tau_{glass}(t) = 1 - \frac{1}{2}\left(\frac{sin^2(\theta_{glass}(t) - \theta_{ARC}(t))}{sin^2(\theta_{glass}(t) + \theta_{ARC}(t))} + \frac{tan^2(\theta_{glass}(t) - \theta_{ARC}(t))}{tan^2(\theta_{glass}(t) + \theta_{ARC}(t))}\right) \tag{3.16}$$

With:

$$\theta_{glass}(t) = arcsin\left(\frac{n_{ARC}}{n_{glass}} sin(\theta_{ARC}(t))\right) \tag{3.17}$$

$n_{air}$ Refractive index of air ($n_{air} = 1$) ($\emptyset$),
$n_{ARC}$ Refractive index of the ARC layer ($n_{ARC} = 1.3$ based on [154, 155]) ($\emptyset$),
$n_{glass}$ Refractive index of glass ($n_{air} = 1.526$ [149]) ($\emptyset$),

Lastly, the effective transmittance through the ARC, $\tau_{cover}$, modules is given by Equation 3.18. When PV plants without ARC are considered, only the transmittance of the glass is investigated (i.e. $\tau_{cover} = \tau_{glass}$).

$$\tau_{cover} = \tau_{ARC} \cdot \tau_{glass} \tag{3.18}$$

The diffuse and ground-reflected solar radiations are considered as isotropic [5], which allows us to integrate the beam transmittance of module's cover over an appropriate range of incidence angles to derive the diffuse and ground-reflected transmittances of the system. Indeed, the transmittance of a system for hemispherical isotropic diffuse radiation can be approximated as the transmittance of the same system for the beam radiation at a specific incidence angle [156]. In [156], the authors performed this integration operation for a variety of configurations and proposed two equivalent angles of incidence to approximate the transmittance of the same system for diffuse and ground-reflected radiation (Equation 3.19). The authors do not stipulate the whole range of configurations tested, but we can assume that ARC was not among the experimental data. This may be neglected, as diffuse and ground-refracted radiation are of second-order influence in comparison with direct radiation.

$$\begin{cases} \theta_g^e = 90 - 0.5788\beta + 0.002693\beta^2 \\ \theta_d^e = 59.38 - 0.1388\beta + 0.001497\beta^2 \end{cases} \tag{3.19}$$

$\theta_g^e$ Equivalent incident angle for ground-reflected radiation (°),
$\theta_d^e$ Equivalent incident angle for diffuse radiation (°).

We would like to emphasise the contradiction between the Perez projection model, which assumes an an-isotropic diffuse irradiance, and the optical model, which considers an isotropic diffuse irradiance. It is assumed that the inaccuracy generated by this simplification is relatively low compared to the errors induced during the forecasting process of PV power.

#### 3.3.3.2 Spectral response

The efficiency of PV cells is sensitive to variations in both the power and spectrum of the incident light.

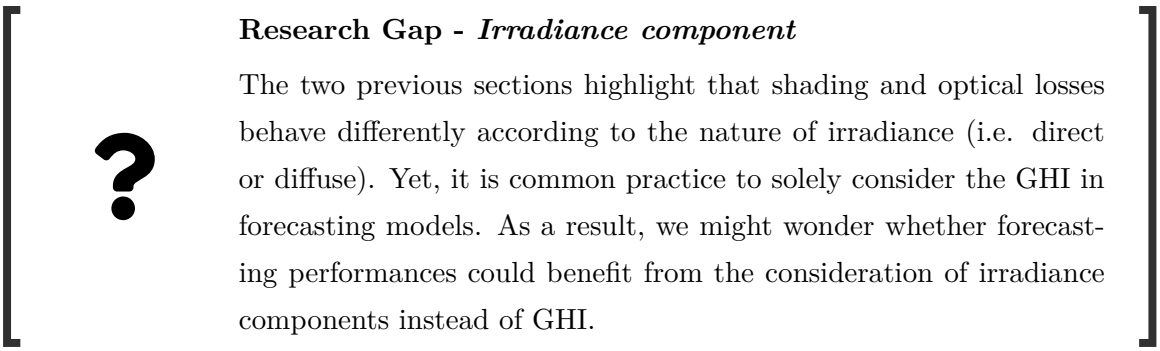
The spectral distribution of light reaching the ground in clear-sky conditions is mainly characterised by the presence of absorption lines due to some molecules, such as ozone, oxygen, and water vapour present in the atmosphere [116]. This distribution also depends on the path length of the ray through the atmosphere: for low solar elevation angles, sunlight passes through a greater proportion of the atmosphere, which leads to a significant Rayleigh scattering of short wavelengths. Besides, clouds alter PV production, not only because they reduce the available downwelling solar irradiance, but also because they act as a spectral filter (water absorbs much more radiation in the near-infrared than in the visible) [148].

Only a few papers try to improve PV power modelling by including the Spectral Response (SR) of modules [148]. This situation can be accounted for by the fact that standard outputs of atmospheric models are broadband data (i.e. integrated over the full short-wave domain), which do not provide any information regarding the spectral distribution. In such

5. This statement can be challenged when considering the circumsolar component of the diffuse radiation (Section 3.3.1.2).

a situation, the spectral effects can be estimated thanks to empirical formulations: [149] uses an air mass modifier which takes into account the Air Mass (AM) that the beam radiation has to cross, while [157, 158] proposes a spectral mismatch factor to account for the fact that the spectral irradiance distribution in the field differs from that of the reference AM1.5 spectrum [6].

The SR of a PV system, i.e. the fraction of available irradiance that is converted into current, is technology-dependent, which means that some technologies can be more or less sensitive to certain bands of the solar irradiance spectrum [159]. For instance, mono-crystalline and multi-crystalline silicone-based modules, which represent more than 90% of the total production [160], are more sensitive to the near-infrared region than to ultraviolet photons [159, 161]. In the PV power modelling literature, the spectral influence upon the conversion performances is below 1% for standard crystalline silicon modules [162, 163] but around 3% for amorphous silicon modules [163]. The spectral mismatch of Crystalline Silicon (c-Si) modules is the lowest among the different PV technologies [115, 158]. The effect of spectral variations can be neglected in this study inasmuch as all PV plants under study are composed of such modules.




**Research Gap - *Irradiance component***

The two previous sections highlight that shading and optical losses behave differently according to the nature of irradiance (i.e. direct or diffuse). Yet, it is common practice to solely consider the GHI in forecasting models. As a result, we might wonder whether forecasting performances could benefit from the consideration of irradiance components instead of GHI.

## 3.4 Conversion of irradiance into electricity

### 3.4.1 Irradiance-to-power modelling

#### 3.4.1.1 Modelling strategies

In the literature, three main approaches are investigated to determine the power output of a PV system.

1. **Equivalent electric circuit-based models:** The current-voltage characteristic of a PV cell varies according to the value of irradiance and the cell's temperature. To assess these characteristics, it is common practice to model the physical behaviour of the cell using an equivalent electric circuit [149] composed of a current source, one or

6. Standard terrestrial solar spectral irradiance distributions, such as the reference Air Mass 1.5 spectrum, can compare the performances of PV devices produced from different technologies or manufacturers. This spectrum is representative of the illumination conditions of the Sun at an elevation angle of about 41° in geographical mid latitudes and under a clear sky.

two parallel diodes and a combination of resistances placed in series and in parallel. The double-diode model is accurate when dealing with low irradiance levels but is outperformed by the single-diode model, which is more relevant for higher illumination conditions. The latter model is more widespread in the literature because it offers a good trade-off between simplicity and accuracy. As the diode model is a nonlinear problem, the parameters estimation is not trivial. Several approaches are presented in literature: (1) analytical solutions can be reached with sets of approximations, (2) iterative methods can be employed, (3) numerical methods are widespread owing to their accuracy and speed of calculation but they may suffer convergence issues due to bad initial conditions [164].

2. **Empirical models:** Another alternative consists in parameterising the physical relation of the PV system [165–167]. Contrary to the previous set of methods, which estimates the power output under varying operating conditions, empirical models assume that the PV cell operates at its MPP. This class of models provides explicit methods and relative easy calculations to obtain the conversion efficiency [168].
3. **Data-based models:** This last class of models avoids the explicit formulation of the physical phenomena that occur during the irradiance conversion process. Instead, the relation is inferred thanks to ML models [168, 169] fed with inputs such as irradiance, ambient temperature and Sun position angles.

This panorama of modelling strategies can be attributed to the areas from which they originated. For instance, laboratories specialised in PV measurements tend to use empirical models, while universities, which are more accustomed to theoretical studies, are the cradle of circuit-based models [170]. These approaches possess advantages and disadvantages. First, both equivalent electric circuit-based models and empirical models emerge from physics-based knowledge. The main distinction lies in the precision of the modelling: the diode modelling strategy is more complex (and usually requires more information) but provides more accurate production estimations [164, 171]. It is worth mentioning that, apart from the parameters estimation of the diode model, the parameters estimation method also plays an important role [115]: improved performances are expected if the estimation is performed on experimental values rather than on values from data sheets. Both studies based their comparison on small installed capacity systems, viz. 2.2 kWp and 9.0 kWp respectively. For such installations, it is relevant to assume that physical parameters and ageing are homogeneous across the various modules. But for large-scale PV plants, such as those under study ($P_c \in (1.3, 12)$ MWp), the variable weather conditions affect each cell differently (e.g. shading effects, small clouds). As a result, the diode modelling strategy seems oversized for our specific case study.

Lastly, data-based modelling is an out-of-the-box solution that does not assume any physical knowledge regarding the conversion process but needs some kind of expert knowledge regarding model tuning. The main limitation to this approach is the need to have

available data for the training process.

In the present work, the empirical model strategy was chosen owing to its low complexity (and the subsequent small programming effort), and its low computational cost.

#### 3.4.1.2 Empirical model

The literature provides us with plenty of power conversion models [166], which aim at integrating the influence of cell temperature over conversion efficiency. Here, we consider three distinct models (Figure 3.6) that we name: (1) power conversion model [165, 172], (2) low irradiance model [165], and (3) efficiency model [173]. These models have been chosen because they require few parameters, which can be found in the technical data sheet [7].

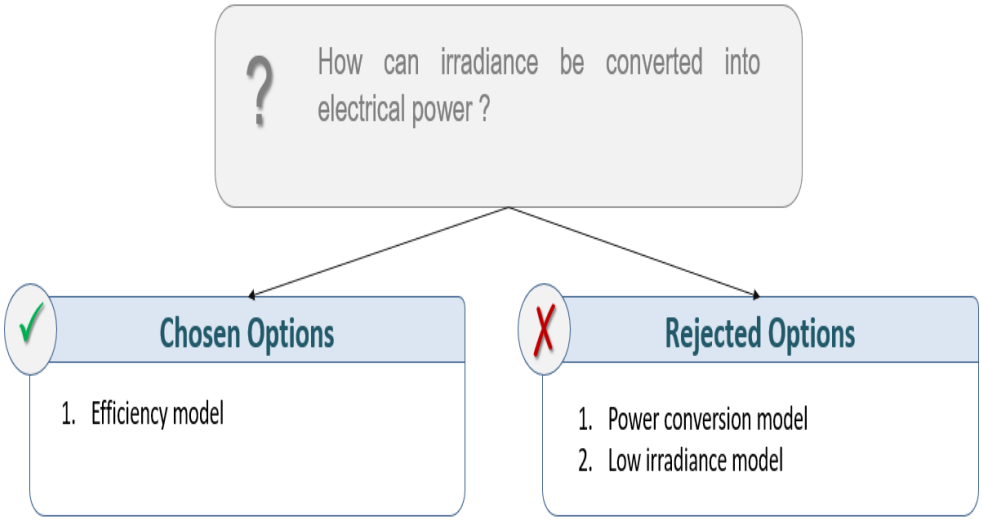



Figure 3.6 – Considered options for the conversion of irradiance into electrical power

The power conversion model (Equation 3.20) is the model most frequently encountered in the literature, probably due to its simplicity and the low number of parameters required. This model supposes a linear dependence between cell temperature and conversion efficiency.

$$P(t) = P_{STC} \cdot \frac{GTI(t)}{GTI_{STC}} \cdot \left[1 + \gamma \cdot \left(T^{cell}(t) - T^{cell}_{STC}\right)\right] \tag{3.20}$$

| | |
|---|---|
| $P(t)$ | Nominal capacity at time $t$ $(W)$, |
| $P_{STC}$ | Nominal capacity at STC $(W)$, |
| $GTI_{STC}$ | GTI at STC $(W/m^2)$, |
| $T^{cell}(t)$ | Cell temperature at time $t$ $(°C)$, |
| $T^{cell}_{STC}$ | Cell temperature at STC $(°C)$, |
| $\gamma$ | Maximum power correction factor for temperature $(°C^{-1})$. |

7. Usually Standard Test Conditions (STC) parameters are: $GTI_{STC} = 1000\ W/m^2$, $AM_0 = 1.5$, $T^{cell}_{STC} = 25°C$

The relationship between the efficiency of a cell and the POA irradiance is nonlinear: as the intensity of incident irradiance decreases, so does the conversion efficiency. This phenomenon is referred to as low irradiance losses [174] and is generally explained by the low shunt resistance of PV modules [175]. Here we consider the two irradiance range models proposed by [165] and defined by Equations 3.21-3.22. The authors highlighted that this approach provided better modelling accuracy than Equation 3.20, yet this approach is not widely used in the literature. Usually, low irradiance losses are neglected by most heuristic models because the dominant contribution to the energy yield comes from higher irradiance [176]. Coefficient $k$ represents the irradiance correction factor and is highly technology-dependent (Table 4 from [165]). For a same family of technology, the spread of $k$ is quite large (e.g. $k \in [0.003 - 0.022]$ for single-crystal Si module).

For $GTI(t) \leq 200\text{W/m}^2$:

$$P(t) = P_{STC} \cdot \left( \frac{GTI(t)}{GTI_{STC}} \cdot \left[ 1 + \gamma \cdot \left( T^{cell}(t) - T^{cell}_{STC} \right) \right] - k \left[ 1 - \left( 1 - \frac{GTI(t)}{200} \right)^4 \right] \right) \tag{3.21}$$

For $GTI(t) > 200\text{W/m}^2$:

$$P(t) = P_{STC} \cdot \left( \frac{GTI(t)}{GTI_{STC}} \cdot \left[ 1 + \gamma \cdot \left( T^{cell}(t) - T^{cell}_{STC} \right) \right] - k \frac{GTI_{STC} - GTI(t)}{GTI_{STC} - 200} \right) \tag{3.22}$$

The efficiency model (Equation 3.23) proposes an alternative where the efficiency is modelled as a linear function of cell temperature under constant air mass and irradiance, but it takes into account nonlinear dependence on air mass and irradiance [173]. p, q, m, r, s, u are technology-dependent parameters (Table 1 from [173]). The production efficiency $\eta$ is then used to derive PV production (Equation 3.24).

$$\eta(t) = p \left[ q \frac{GTI(t)}{GTI_{STC}} + \left( \frac{GTI(t)}{GTI_{STC}} \right)^m \right] \times \left[ 1 + r \frac{T^{cell}(t)}{T^{cell}_{STC}} + s \frac{AM(t)}{AM_0} + \left( \frac{AM(t)}{AM_0} \right)^u \right] \tag{3.23}$$

$$P(t) = \eta(t) \cdot A \cdot GTI(t) \tag{3.24}$$

| | |
|---|---|
| $A$ | Active cell area of the module $(m)$, |
| $AM(t)$ | Relative air mass (Equation 3.8) $(\emptyset)$, |

For the sake of brevity, we do not include a forecasting performance comparison of these three power conversion models. Within the frame of this thesis, it has been observed that the efficiency model outperforms other modelling strategies.

### 3.4.2 Operating temperature

When irradiation reaches the PV cell, some of the energy is converted into electricity while the remaining part becomes heat, which significantly increases the temperature of

the module and reduces its efficiency [167]. Actually, the increase in temperature causes a narrowing of the band-gap energy (i.e. the minimum amount of energy required for an electron to break free from its bound state), which induces a higher generation of electron-hole pairs and increases the short-circuit current. However, at the same time, the higher temperature also decreases the PV voltage [177]. Ultimately, the increase in current is not enough to offset the decrease in voltage, which results in a power drop. Apart from the absorbed incoming irradiance, the PV cell's temperature is also influenced by ambient climatic conditions (e.g. temperature and wind) as well as the module technology (e.g. cell technology, optical properties of cover). Wind reduces the operating temperature thanks to forced convection: Fig.5 from [178] shows a wind cooling effect of around 17°C for wind speeds of $9 - 10\ m/s$ at a 1000 $W/m^2$ global irradiance.

Thus, the influence of temperature cannot be ignored: the technical data sheet of the 60P250-Sillia poly-crystalline module [179] reports a 0.42% drop in maximum power produced for each degree rise in temperature w.r.t. STC. A variation in cell temperature of 40°C results in a 16.8% drop in the maximum power (at 65°C the module's maximum power drops from 250 Wp to 208 Wp). During summer clear-sky days, a cell's temperature can easily reach 60°C for free-standing systems in central Europe [180].

A comparison between several models estimating cell temperature highlights that [180]:

- the inclusion of the cooling effect resulting from wind provides better estimations of the cell's temperature,
- wind data from an NWPs model allows an estimation of the module temperature with an error of the same order of magnitude as in-situ data.

Following the conclusions of this article, we consider the module temperature model introduced in [181] and defined by Equations 3.25-3.26.

$$T^{cell}(t) = \frac{U_{PV}(w(t)) \cdot T^a(t) + GTI(t) \cdot [\tau_{cover} \cdot \alpha_{cell} - \eta_{STC}\,(1 + \beta_{STC} \cdot T_{STC})]}{U_{PV}(w(t)) - \beta_{STC} \cdot \eta_{STC} \cdot GTI(t)} \tag{3.25}$$

With:

$$U_{PV}(w(t)) = 26.6 + 2.3 \cdot w(t) \tag{3.26}$$

| | |
|---|---|
| $U_{pv}(w)$ | Heat exchange coefficient for the total surface of the module, |
| $T^a(t)$ | Ambient temperature at time $t$ ($°C$), |
| $\tau_{cover}$ | Transmittance of the cover system, |
| $\alpha_{cell}$ | Absorption coefficient of the cells, |
| $\eta_{STC}$ | Efficiency coefficient of maximal power under STC, |
| $\beta_{STC}$ | Temperature coefficient of maximal power under STC ($°C^{-1}$), |
| $w(t)$ | Local wind speed close to the module ($m/s$). |

The authors used $\tau_{cover} \cdot \alpha_{cell} = 0.81$. Lastly, the wind speed at ground level is obtained via the power law:

$$w(t) = w_{10}(t) \cdot \left(\frac{Z}{10}\right)^{\left(\frac{1}{7}\right)} \tag{3.27}$$

$w_{10}(t)$ Wind amplitude at 10m height $(m/s)$,
$Z$ Panel height $(m)$.

### 3.4.3 Ageing

A PV system is generally expected to be in operation for at least 20 years. Over time, PV modules and electric conversion components (e.g. inverters) degrade under actual operating conditions, which reduces their energy yield. Cell performance degradation mechanisms are mainly material ageing, corrosion, metal mitigation through the p-n junction, cracked cells, bypass diode failures, and changes in the module series resistance [182]. In addition, external factors such as shading (which induces hot spots) and high temperature increase the energy-yield degradation rates of modules [183]. The authors highlight that the degradation of modules occurs at approximately $-0.5\%/year$, which is in line with the findings of [162], which proposes an average annual decrease in performance of approximately 0.8%.

As the influence of ageing is negligible, this process has not been modelled within the physics-based conversion architecture. However, for further studies, a linear degradation of the conversion efficiency according to the time can be considered. In the context of a statistics-based model, the notion of flow of time can be easily integrated in the regression model by considering a proxy of time.

### 3.4.4 Electricity conversion

The last stage consists in converting electricity to comply with grid injection requirements. Power inverters are used to convert Direct Current (DC) generated by modules to AC and synchronise it with the grid voltage, while the transformer increases the voltage before interfacing with the electrical grid.

In general, the efficiency of a PV inverter is a function of the input power and input voltage. At medium to high levels of irradiance, the inverter has a high efficiency, typically greater than 90%, while at very low irradiance levels, the efficiency drops sharply [184]. Fig 1.4 from [184] shows that inverter efficiency can be considered as constant over a wide range of output power (namely for output power higher than 20% of rated power). An investigation study performed in [185] examined several approaches to model the efficiency of grid-connected inverters. Three inverter efficiency models (including the one proposed in [185]) are investigated in [115]. These models are considered within a physics-based conversion chain of irradiance into AC power. The results indicate that inverter models

have a minor impact on the overall performance of the modelling chain. Therefore, the impact of the inverter efficiency is neglected in this work.

**Research Gap - *Accuracy***

Physics-based models are generally chosen due to their easy implementation and reduced computing efforts. Such requirements lead to model simplifications that can induce a lack of accuracy. In addition, several processes in the conversion chain have been neglected, while parameters such as site geometry may be spurious (Section 4.2). Thus, it is legitimate to wonder whether a regression model fed with preprocessed irradiance would outperform the same model trained on a relevant set of inputs accounting for the physics behind the conversion chain. The objective is to determine whether a statistical model, provided with the relevant features, is able to derive irradiance conversion laws in addition to extrapolating weather variations. In other words, can we improve forecast accuracy by converting irradiance inputs into power-like feature?

## 3.5 Evaluation

The previous section enables us to get an insight into the various processes involved in the conversion of irradiance into AC power. A detailed summary of the selected models and the resulting architecture is presented in Figure 3.1. Figure 3.7 gives a simplified representation of the conversion chain. In this section we focus on the projection, shading, optical and efficiency models respectively defined in Sections 3.3.1.2, 3.3.2, 3.3.3, 3.4.1.2.

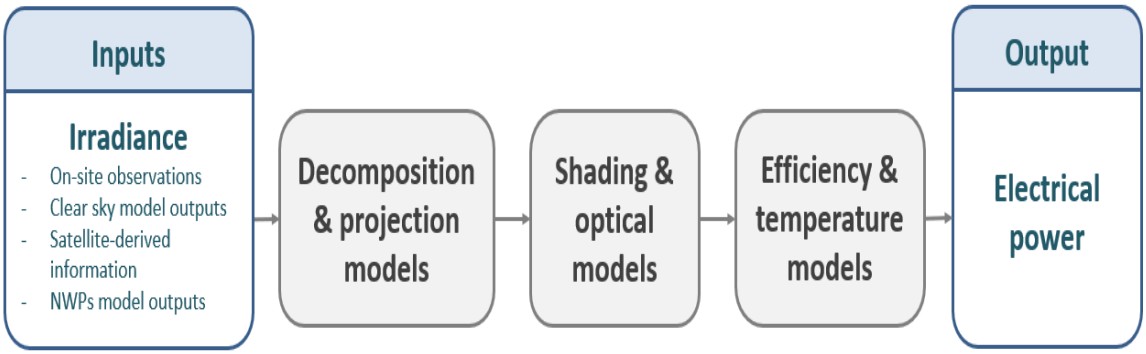



Figure 3.7 – Schematic representation of the conversion chain.

### 3.5.1 Explicit conversion of irradiance

In this section we investigate the impact of the proposed conversion architecture over the properties of irradiance features as well as the impact on forecasting performances.

#### 3.5.1.1 Linear dependency on production

Figure 3.8 represents the evolution of the linear relationship existing between pre-processed irradiance (in the absence of on-site GHI observations, satellite-based observations and NWPs model outputs are examined) and observed production. The different pre-processing steps consist in the projection of GHI onto the POA (and the consideration of shading and optical effects) and in the conversion of GTI into an electrical power-like feature. First, Figure 3.8a exhibits a clear linear dependency between the GHI and the observed production, and a significant spread of points. The projection model introduced in Section 3.3.1.2 increases the concentration of points along the diagonal, while improving the coefficient of determination (Figure 3.8c). The power conversion model has little influence over the coefficient of determination. However, we observe in Figure 3.8e that the latter corrects the deviation due to temperature for high levels of irradiance. Similar conclusions are reached when considering forecast irradiance instead of satellite-derived irradiance (Figures 3.8b, 3.8d, 3.8f). Overall, scatter plots obtained with irradiance forecasts are broader. The spread of the scatter plot is due to the coarse spatial resolution of the numeric weather model as well as inherent forecasting errors. The collinearity degree of data is mainly due to the projection model, while few contributions are observed from the shading and optical modelling (which are not shown for the sake of clarity), and the irradiance-to-electricity conversion.

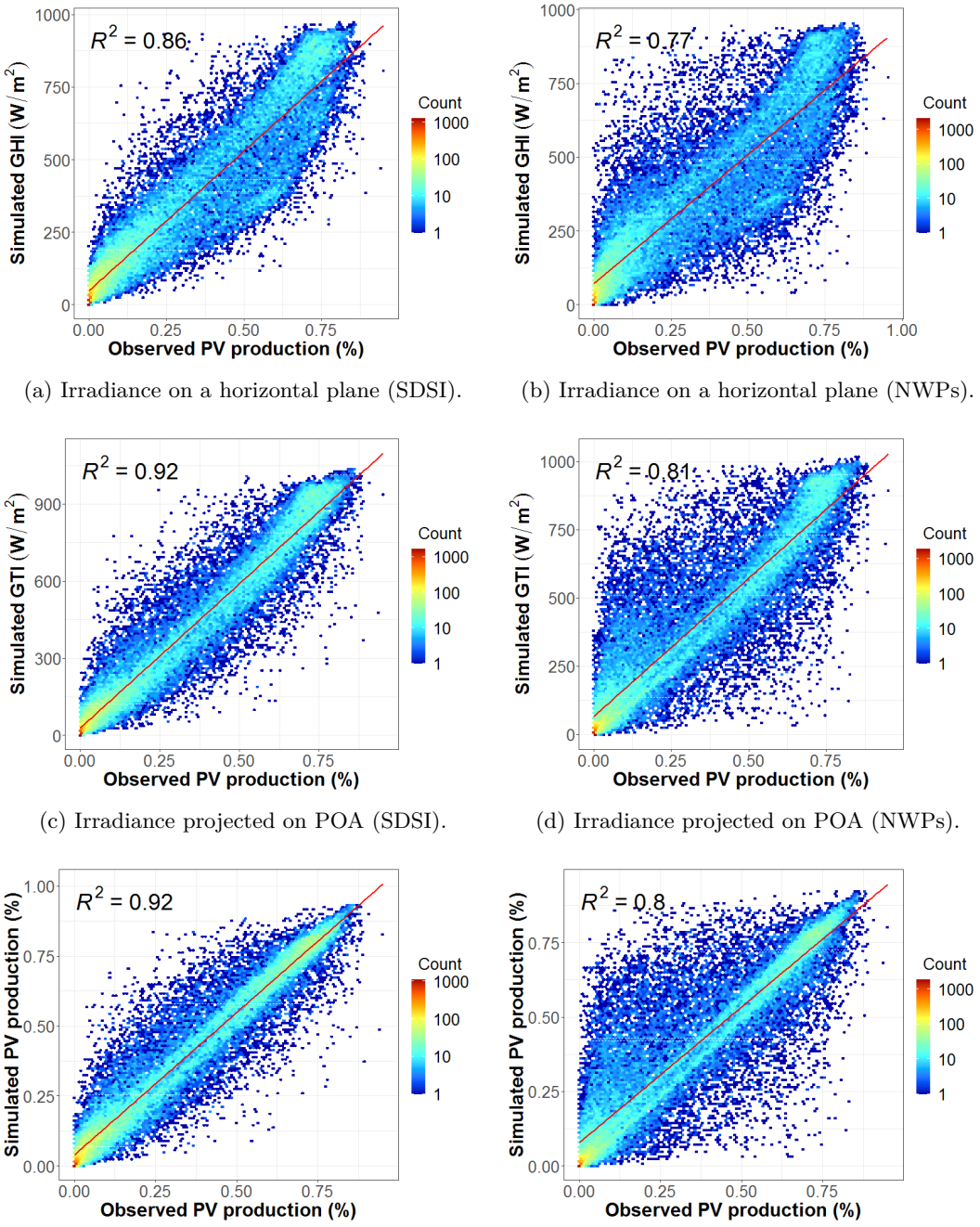
Consequently, the proposed modelling chain strengthens the linear relationship between the response and explanatory features. Such characteristics may be sought when on-site measurements are used to assess the quality of production measurements.

#### 3.5.1.2 Impact on forecasting performances

In the ML-related literature, it is common practice to resort to features (namely GHI) which have not been normalised by clear-sky model outputs. Therefore, it is insightful to analyse the impact of the different elements of the irradiance-to-electricity conversion chain on forecast irradiance in the context of PVPF.

Figure 3.9 represents the forecasting performances of the RF model fed with past production observations and features derived from irradiance forecasts. This figure clearly demonstrates that in the case of non-normalised inputs, the projection of forecast irradiance over the POA (i.e. *RF + GTI(P)* model) improves forecasting performances for the three scores under study. The inclusion of shading and optical effects (i.e. irradiance reflection) slightly improve the normalised Root Mean Square Error (nRMSE) and normalised Mean Absolute Error (nMAE) scores. However, the consideration of the efficiency variations slightly degrades the scores.

Given that the forecasting performances of the models fed with pre-processed forecast irradiance are very close, it is insightful to look at the statistical significance of forecast differences. Figure 3.10 represents the Diebold-Mariano (DM) statistic between *RF+GTI(P)*,




(a) Irradiance on a horizontal plane (SDSI).

(b) Irradiance on a horizontal plane (NWPs).

(c) Irradiance projected on POA (SDSI).

(d) Irradiance projected on POA (NWPs).

(e) Production derived from irradiance (SDSI).

(f) Production derived from irradiance (NWPs).

Figure 3.8 – Binned scatter plots of observed production and preprocessed irradiance. The figures on the left are obtained considering SDSI observations, while those on the right result from forecast irradiance at (PV6) for the years 2015 and 2016. The red line represents the regression line obtained with the least-squares method.

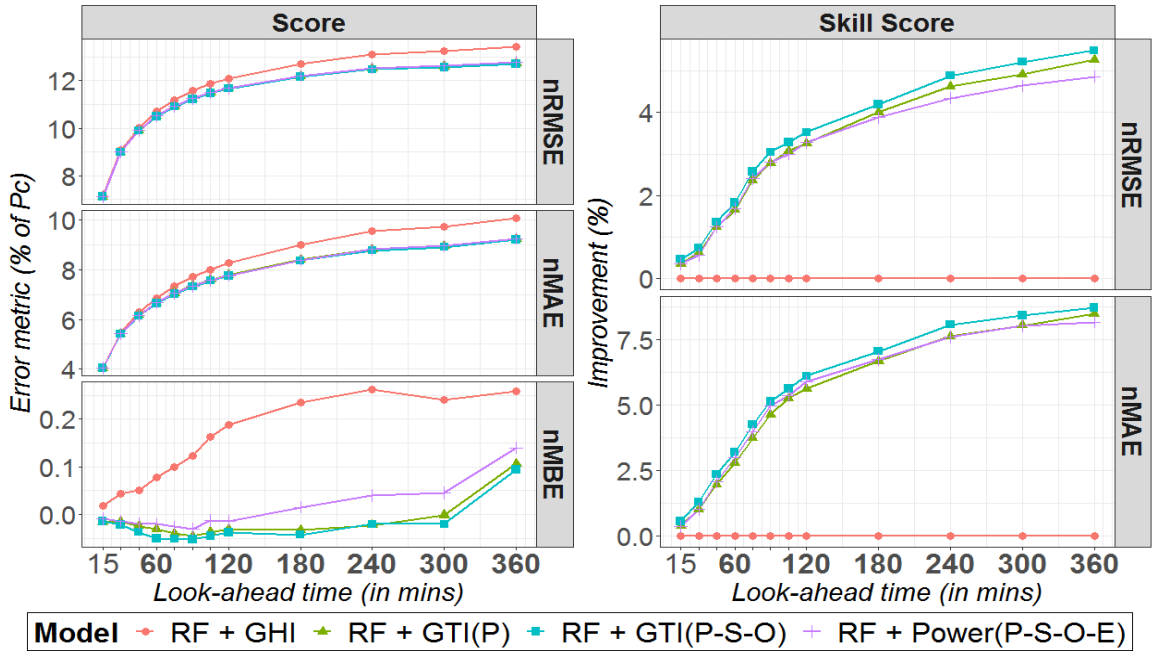



Figure 3.9 – Performances obtained with Random Forest (RF) fed with non-normalised inputs (i.e. past PV observations and irradiance forecasts). The influence of the projection (P), shading (S), optical (O), and efficiency (E) models are analysed.

*RF+GTI(P-S-O)*, and *RF+GTI(P-S-O-E)* models. The graph shows that the difference between forecasts produced with the *RF+GTI(P-S-O)* and *RF+GTI(P)* models are statistically significant. This highlights the relevance of the shading and optical models.

**Research Answer - *GHI conversion***

Section 3.5.1.2 shows that the conversion of irradiance has a positive impact on forecasting performances. In the case of non-clear-sky normalised forecast irradiance, the major source of improvement results from the projection model, which is in line with findings in [115].

### 3.5.2 Implicit derivation of physical laws

In Sections 3.3 and 3.4, we developed a physics-based conversion chain based on primitive features (e.g. irradiance components, solar angles, temperature), which generates AC power from irradiance data. In Section 3.5.1, we highlighted that the explicit conversion of GHI into products including plant characteristics achieves higher accuracy in the context of PVPF. At this point, we seek to determine whether statistical algorithms are able, to some extent, to implicitly derive conversion laws from initial features rather than post-processed ones. In other words, our aim is to investigate the optimal set of inputs with which to provide the model to obtain the highest forecast accuracy.

First, we investigate the influence of forecast radiation components (i.e. BHI and DHI) on the skill of the PVPF model. We opt for the RF model and its nonlinear capabilities.

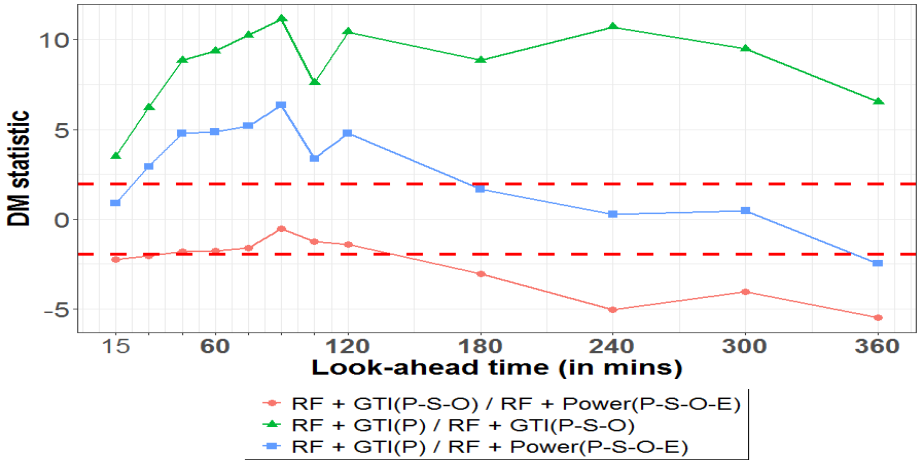



Figure 3.10 – DM statistic (defined in Section 2.3.3) between the RF models outputs fed with non-normalised features derived from irradiance forecasts. The influence of the projection (P), shading (S), optical (O), and efficiency (E) models are analysed. The red dotted lines show the borders delimiting the validation and rejection of the null hypothesis.

In Figure 3.11, we observe that the forecasting performances achieved by the model fed with irradiance components outperform those of the model based solely on GHI in terms of nRMSE, nMAE and normalised Mean Bias Error (nMBE). Therefore, knowledge of the diffuse and direct parts of irradiance allows a better characterisation of the atmosphere and consequently a better inference of production laws. However in the end, these improvements are eclipsed by the consideration of the power-like feature within the *RF + Power(P-S-O-E)* model.

**Research Answer - *Irradiance components***

In the case of the nonlinear model, the consideration of the irradiance components improves nRMSE and nMAE scores w.r.t. to a similar approach considering GHI. Nonetheless, the best performances are reached when considering the power-like feature or simply the GTI (Figure 3.9 depicts small performance differences between the RF models fed with these inputs).

Second, we investigate the ability of the RF to work with the initial features used to model the irradiance-to-electricity conversion laws. To do so, we compare the performances of the model fed with the primitive features w.r.t. the same model fed with pre-processed irradiance. We consider the following sets of initial features [8]:

1. Features set 1: DHI, BHI, AM, $\theta$, $\alpha^S$, $\theta^S$,
2. Features set 2: DHI, BHI, AM, $\theta$, $\alpha^S$, $\theta^S$, temperature, and wind amplitude.

The first set is related to the projection, shading and optical effects, while the second takes into account the thermal dependence of the efficiency. We observe in Figure 3.12 that

8. The incidence angle is added because it carries information regarding the plant's geometry.

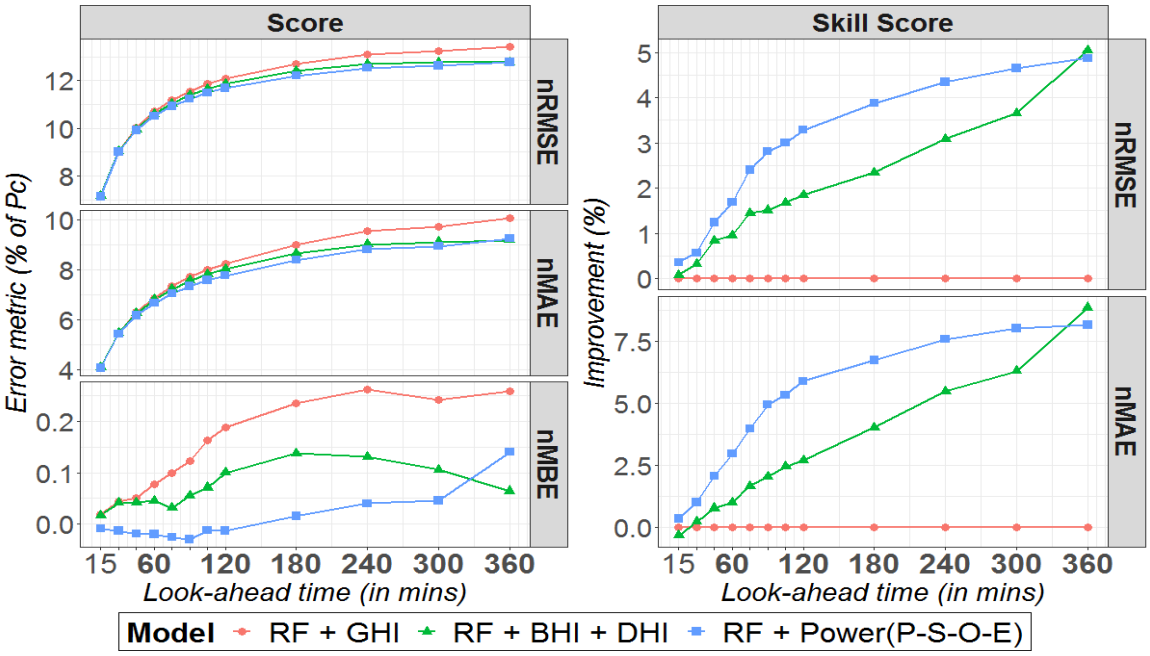



Figure 3.11 – Forecasting performances of the RF models fed with GHI, BHI-DHI forecast, and pre-processed irradiance (i.e. *RF + Power(P-S-O-E)*). To highlight performance variations, the reference model is set as *RF + GHI*.

the RF model fed with *features set 1* outperforms the forecasts derived from the physics-based conversion of irradiance (i.e. model *RF + Power(P-S-O-E)*) both in terms of nRMSE and nMAE. This is associated with an increase in bias, but this is of secondary order due to the very low values observed. A hasty conclusion would be to think that the RF is better at modelling the conversion process than the methodology that we propose. However, we have to keep in mind that the RF deals with several processes, including the prediction and the conversion processes. Within the scope of prediction, solar angles play a significant role, all the more so as we consider non-normalised clear-sky features in this section. This analysis is corroborated by the fact that the model *RF + Power(P-S-O-E) + Solar angles* outperforms the model *RF + Power(P-S-O-E)*.

Moreover, it is possible to slightly extent forecast accuracy (at least in terms of nMAE) by considering temperature and wind amplitude. Given the low accuracy improvement, the inclusion of features related to thermal effects is of secondary order compared to other features such as solar angles. This observation is corroborated by Figure 3.13, which represents the features importance of *RF + Features set 2* model. The graph highlights that main features are irradiance components and solar angles. This low importance of temperature may be explained, to some extent, by the fact that irradiance and temperature are correlated. Therefore, we can assume that the regression model is able to implicitly account for thermal-based efficiency drop thanks to the knowledge of irradiance levels reaching the panels.

Based on these results, it is difficult to point out which methods is preferable considering

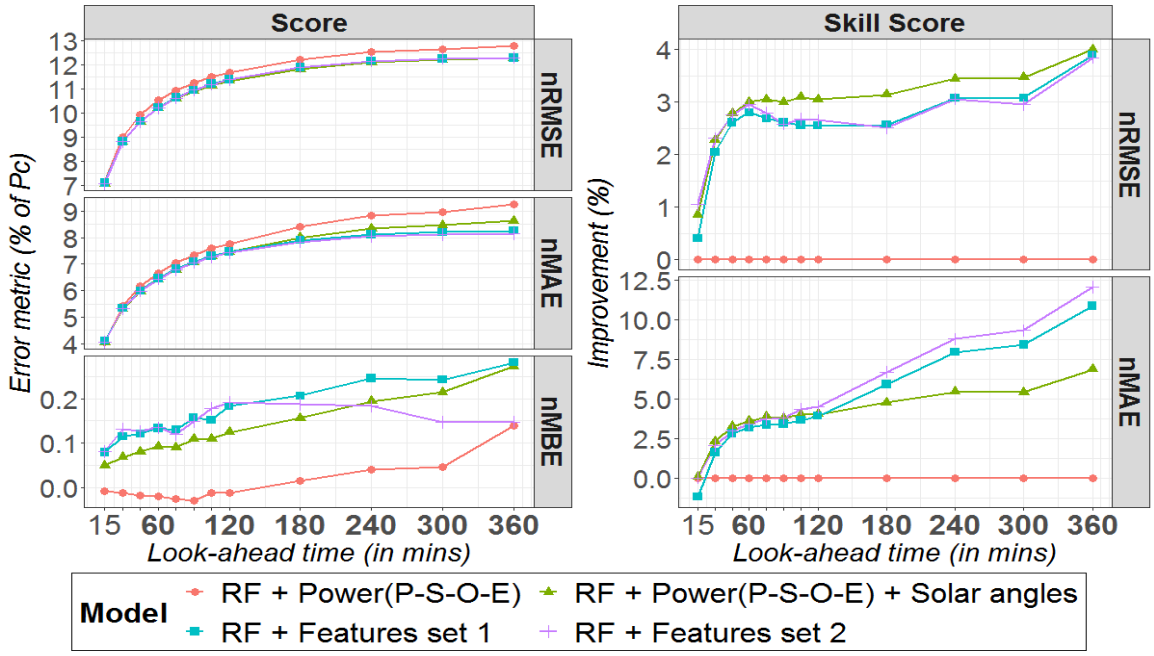



Figure 3.12 – Forecasting performances of the RF models fed with pre-processed forecast irradiance and solar angles (i.e. *RF + Power(P-S-O-E) + Solar angles*) as well as sets of initial features used to derive electrical production from GHI. To highlight performance variations, the reference model is set as *RF + Power(P-S-O-E)*.

(1) pre-processed features, or (2) initial features. In terms of accuracy, the former leads to slightly better nRMSE scores, while the latter provides a significant improvement for the nMAE metric. The second approach possesses the advantage of a low modelling complexity inasmuch as solar angles are easily retrievable from dedicated packages. Given the close accuracy of forecasts obtained with initial or pre-processed features, we can assume that nonlinear regression tools such as RF are able to implicitly derive physical conversion laws from the features sets considered. We may hypothesise, for instance, that when knowing the irradiance levels and solar angles, the model is able to divide its search space into sub-spaces for which modules have experienced a similar spectral response.

**Research Answer - *Accuracy***

In the case of nonlinear forecasting models, the use of physics-based models to convert irradiance into electrical power improves forecasting performances by explicitly integrating the physical knowledge of the processes at stake. However, we highlighted that a statistical model fed with a relevant set of features intervening in the conversion chain also leads to a significant performance improvement. Based on forecast accuracy it is difficult to determine which option is the best, but the latter option does not need explicit modelling.

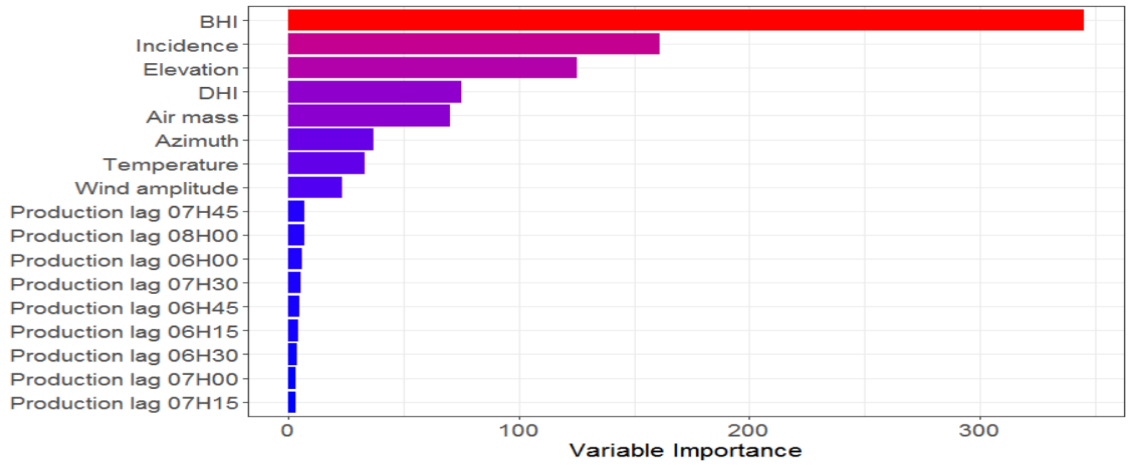



Figure 3.13 – RF impurity-based feature importance (defined in Section 2.2.4). The model generates forecasts for a 6-hour ahead horizon at plant PV1. The low importance of production observations is due to the fact that for such horizons the main source of information is NWPs.

## 3.6 Conclusions

The work performed in this chapter provides an in-depth description of physical phenomena occurring during the conversion of solar irradiance into AC power. From a knowledge and information perspective, this study assesses the models available in the literature and summarises the main factors impacting PV production. The methods developed can be viewed as a way to inject physics-based information into the regression models.

The conversion of irradiance into electrical power is modelled through a modular chain of several sub-models dedicated to specific aspects of the conversion. In the context of PVPF, the conversion model is applied to irradiance information to derive plant-specific features (such as GTI or power-like variables). This kind of physics-based approach is insightful for newly built power plants without production records. In the case of ML models, we highlighted that an explicit integration of physical knowledge through the pre-processing of irradiance leads to higher forecast accuracy w.r.t. a straightforward integration of the GHI. In addition, the description of the conversion processes highlighted that the irradiance components are affected differently according to the physical process at stake. This inspired the use of DHI and BHI instead of the global irradiance. The former approach turns out to be more efficient in terms of forecast accuracy. In the same vein, we shown that directly using the primitive inputs (namely inputs intervening within the conversion steps) to fit an RF model provides somewhat comparable forecasting performances to the output of the physics-based modelling. Therefore, ML tools such as RF are able to reckon conversion laws when fed with relevant information. In both cases, the projection model is the critical one, which is in line with [115]. The consideration of thermal effects is of secondary order compared to the influence of solar angles.

In line with the quality analysis performed in Chapter 4, physics-based models are vital

to convert on-site or off-site observations before using them as proxies of production performances. The work performed in this chapter is continued in the next chapter. More precisely, the impact of the conversion chain is analysed in the context of clear-sky normalisation in Section 4.4.

## 3.7 Résumé en Français

Pour générer des prévisions de la production PV, plusieurs approches s'offrent à nous. Il est possible d'utiliser, comme nous l'avons vu dans le précédent chapitre, des modèles de régression statistique. Dans ce cas, le modèle se charge d'inférer la relation entre les variables d'entrée et la variable de sortie et agit comme une boîte noire ou grise, suivant la complexité du modèle utilisé. L'alternative consiste à modéliser explicitement la chaîne de conversion de l'irradiance en électricité via les lois physiques appropriées. La conversion de l'irradiance en courant électrique alternatif est un processus complexe mobilisant de nombreux composants et dépendant de facteurs externes, le meilleur exemple étant l'impact de la géométrie (i.e. l'orientation et l'inclinaison des modules) sur la production. En soit, cette famille de modèle ne réalise pas des prévisions puisque cette tâche est dévolue aux modèles NWPs en charge de prévoir l'irradiance et les variables entrant dans la chaîne de conversion. Les modèles physiques sont très peu utilisés dans la littérature traitant de la prévision de la production PV. Ce phénomène peut s'expliquer par le manque de données techniques concernant l'architecture et la configuration des centrales PV.

Les objectifs de ce chapitre sont doubles : on cherche d'une part à modéliser la conversion de l'irradiance en puissance électrique via un ensemble de formules simples et ne nécessitant que très peu de paramètres, et d'autre part à intégrer ces connaissances dérivées de la physique dans le modèle de prévision statistique dans l'optique de réduire l'effort de calcul de ce dernier et d'accroître sa précision. Afin de guider le lecteur, la Figure 3.1 nous fournit une vision synoptique de la chaîne de modélisation.

### Altération de l'irradiance

La première étape de cette chaîne consiste à modéliser les altérations que subit l'irradiance avant d'atteindre la cellule PV.

Généralement, les modèles NWPs, les modèles ciel clair ou tout simplement les mesures réalisées par un pyranomètre nous fournissent la valeur de l'irradiance globale sur une surface horizontale au niveau du sol. Il est donc d'abord nécessaire de projeter cette quantité sur le plan des modules. Pour ce faire, il nous faut connaitre la valeur des trois composantes de l'irradiation globale horizontale, à savoir la composante diffuse, directe, et celle reflétée par le sol. Ces informations peuvent être directement fournies par les modèles ou certains appareils de mesure. Dans le cas contraire il est nécessaire de recourir à un modèle de décomposition. Dès lors, il est possible de projeter chaque composante via un modèle dédié.

La projection de la composante diffuse se décline sous une multitude de modèles en raison de la complexité de sa modélisation, alors que la projection des deux autres composantes est plutôt éprouvée.

La connaissance des trois composantes permet également de modéliser l'ombrage inter-rangées. Les ombres portées par les structures alentours telles que les constructions humaines ou la végétation ne sont pas prises en compte. Par contre, un masque est appliqué pour prendre en considération les reliefs topographiques locaux.

Nous avons également tenté de prendre en compte l'effet de l'encrassement des modules. Le seul nettoyage que subissent ces derniers est celui apporté par les pluies. Or, à partir des données à disposition, nous n'avons pas été en mesure de mettre en évidence d'éventuelles variations de la puissance avant et après les précipitations. Ceci est supposé résulter de diverses erreurs de mesures et de modélisation qui sont supérieures à la baisse de performance induite par l'encrassement des panneaux. Enfin, les effets optiques sont intégrés à la chaîne de modélisation.

Un module PV est composé de différentes couches qui réfléchissent et absorbent l'irradiance différemment. En ce qui concerne la composante directe, ces phénomènes sont modélisés via les équations de Fresnel et de Snell. Une modélisation spécifique est apportée pour les deux autres composantes et repose sur des données empiriques. L'efficacité des cellules PV est tributaire du niveau d'irradiance mais également du spectre de la lumière incidente. Dans la mesure où la majorité des sites étudiés est constituée de modules silicone cristallin (c-Si), matériau relativement peu sensible au décalage spectral, nous avons fait le choix de négliger cet effet.

### Conversion de l'irradiance en électricité

La seconde étape de ce processus de conversion réside en la conversion de l'irradiance en électricité et en sa transformation avant injection sur le réseau. Tout d'abord, les cellules PV permettent de convertir la lumière incidente en courant continu. Cette étape peut être modélisée selon trois approches distinctes : via (1) un modèle électrique équivalent (i.e. modèle diode), (2) un modèle empirique ou (3) une modélisation statistique (e.g. modèle de ML). Dans la mesure où cette dernière option est gourmande en données, et que la première modélisation est relativement complexe, nous avons fait le choix d'opter pour l'approche empirique. Ce type de modèle requiert en entrée des paramètres standard propres à la cellule considérée que l'on peut facilement trouver sur les fiches techniques des modules, ainsi que l'irradiance incidente et la température de la cellule. En effet, plus la température de la cellule est importante et moins le rendement de conversion est bon. Cette influence est modélisée en prenant en compte des paramètres soit empiriques, soit spécifiques à la technologie employée, ainsi que la température extérieure et le vent à proximité du sol.

Le processus de conversion du courant continu en courant alternatif est laissé à la discrétion du modèle statistique dans la mesure où la littérature met en avant la faible influence

des rendements de conversion des onduleurs et transformateurs vis-à-vis du reste de la chaîne de conversion.

### Impact de la modélisation physique sur les performances prédictives

Notre modèle de conversion est maintenant achevé. La suite de ce chapitre consiste à en évaluer l'influence sur les modèles de prévision.

Dans le domaine de la prévision de la production PV, et plus spécifiquement le pan de littérature traitant des modèles d'apprentissage machine, il est courant de considérer directement l'irradiance globale sur plan horizontal sans aucune étape de prétraitement. Nous mettons en évidence que la projection de l'irradiance globale sur plan incliné permet d'améliorer les performances du modèle RF, et que la prise en compte de l'ombrage et des effets optiques a également une influence positive, mais moindre, sur la précision.

A ce stade, nous avons cherché à savoir si le modèle de régression était capable d'intuiter les relations de conversion en jeu. Pour ce faire, nous lui avons fourni en entrée les variables « élémentaires » (e.g. la composante directe et diffuse de l'irradiance, les angles solaires). Il se trouve que lorsque le modèle RF est alimenté par ces variables élémentaires, il atteint des niveaux de précision supérieurs à ceux obtenus en considérant l'information prétraitée par la chaîne de modélisation définie précédemment. Ceci prouve la capacité du modèle non-linéaire à déterminer implicitement les relations de conversion. Il est toutefois possible d'améliorer les performances du modèle RF calé sur la puissance électrique dérivée du modèle de conversion en adjoignant les angles solaires.

# Chapter 4

# Data Characterisation

To believe with certainty we must begin with doubting.

Stanislas Leszczynski (1764)

## Contents

## 4.1 Introduction

Today, data may be considered as the new oil of our 21$^{th}$ century economy [186, 187]. Both are raw materials that, combined with relevant technologies, provide the foundations of specific products and services that we resort to in our daily life. For instance, fuel, plastic and pills are derived from petrochemicals, while autonomous cars, industrial automation, and product recommendations are made possible thanks to Artificial Intelligence (AI)-driven technology [1]. The latter involves Machine Learning (ML) algorithms, which can learn some specific tasks from historical databases without being explicitly programmed. Therefore, this emerging digital economy relies on data-driven technologies. Most of the time, data cannot be simply fed into ML algorithms but rather have to be pre-processed to extract their valuable product: *information*. This process can be viewed as a kind of *data refinement*, which involves expert skills to exhibit relevant patterns, while ensuring data integrity.

### 4.1.1 Objectives

This chapter aims at assessing the characteristics of the different sources of information that we have at our disposal and to define the limits and potential of their applications in the scope of Photovoltaic Production Forecasting (PVPF). Irradiance-related sources of information are complex signals composed of distinct components. Examples include (1) the deterministic influence of the Sun over seasonal and daily variation patterns, (2) the stochastic impact of weather structures (e.g. clouds) and, in the case of power production (3) variations due to the failures of production components. This raises several challenges:

1. First, inasmuch as power observations result from real-world measurements, they may be incomplete or corrupted. The very first challenge lies in identifying these fallacious data and providing appropriate treatment.

2. Second, dependencies between the different inputs can be spoiled because of these several sources of variability. Indeed, in the context of Spatio-temporal (ST) forecasting, stations aligned on the east-west axis could exhibit high correlation scores because of the Sun's path rather than effective cloud dependencies. Therefore, the second challenge tackled in this chapter is related to the expression of the relevant information contained within the inputs.

In a nutshell, the main objective of this chapter is to define the characteristics of inputs and highlight their dependencies by eliminating sources of variability that are not directly related with stochastic weather phenomena impacting Photovoltaic (PV) production.

1. Artificial intelligence can be defined as the field of computer science that tries to mimic human intelligence.

### 4.1.2 Methodology

In this thesis, we adopt a multi-source approach to extend the forecasting performances of traditional models. As a reminder, Section 2.4 details characteristics of production observations, satellite-based observations and Numerical Weather Predictions (NWPs) model data. Section 4.2 proposes an imputation strategy based on a clustering algorithm to deal with the corrupted production observations. Then, in Section 4.4 temporal and spatio-temporal correlations observed within production time series and with spatially distributed production observations or satellite-based datasets are investigated. The general workflow of this chapter is displayed in Figure 4.1.

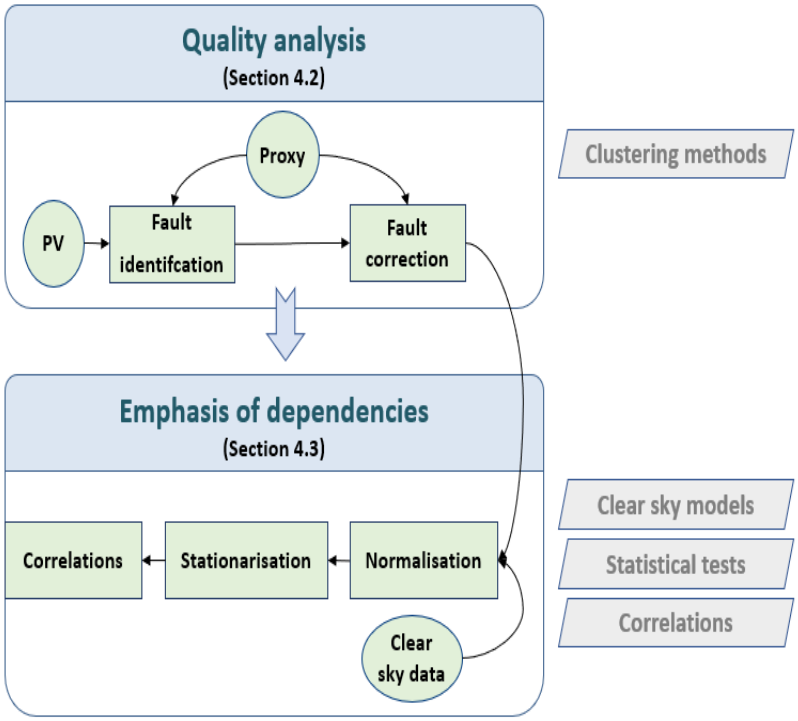



Figure 4.1 – General workflow of the chapter.

## 4.2 Data quality analysis

At this stage, it becomes relevant to question the integrity of available data. Our aim here is to perform a critical analysis to identify data that may deviate from an objective characterisation of reality. This section is motivated by the fact that manually written log files summarising production failures are hardly usable due to a lack of precision regarding event type and temporal occurrence.

### 4.2.1 Metadata reliability

The first step of our investigation consisted in reassessing the system metadata: namely the tilt and azimuth angles of the modules. These angles directly impact the projection of

irradiance on the plane of array. Under our latitudes, $\beta = 25°$ and $\alpha = 180°$ are typical values used during project sizing but reality on the ground can be somewhat different: e.g. panel orientation may be imposed by local topography (Figure 4.2).

Figure 4.2 – Photograph of PV2 [188].

A basic approach was investigated at a very early stage of the thesis to determine a set of angles $(\alpha, \beta)$ in line with the actual site configuration. These angles are used in the forecasting process, and more precisely during the normalisation step of the PV production (i.e. module 2 of Figure 1.10), which involves the projection of the clear-sky irradiance onto the Plane-of-Array (POA) (Section 3.3.1.2). The main idea behind this approach was to assume that the set of optimal angles would provide the most accurate forecasts. This led us to a grid-search algorithm aimed at optimising the forecast accuracy through the orientation angle values. Ultimately, the computed angles were not retained owing to some wide deviations from standard values and the weakness of the proposed approach. Thus, the angles from the technical specifications were kept.

Retrospectively, other methods could have been employed. For instance, it is possible to filter out a set of observed clear-sky production curves and to select angles that lead to the best fit with the modelled clear-sky production curves. This idea is implemented in [189], where the authors apply a nonlinear least-squares solver to derive orientation angles and the loss factor from the equation $P_{meas,cs}\ P_{sim,cs}(\alpha, \beta, LF)$ (an average accuracy in terms of the system orientation of 4° is achieved). In our case, such an approach could be used with the physics-based modelling described in Chapter 3, but it was not implemented due to time constraints. In the context of this thesis, we must consider that some parks have undersized inverters, which impacts the shape of the clear-sky production profile.

### 4.2.2 Production observation reliability

PV power plants are complex systems involving several physical processes to convert solar irradiance into Alternative Current (AC). The sources of variability associated with the output energy result mainly from the weather conditions, but they may also be impacted by technical failures. The latter can induce information losses in the form of missing or erroneous observations which do not reflect reality. In such conditions, it becomes challenging to establish relevant statistical models dedicated to forecast generation when exogenous

sources of data are used.

In the PVPF literature it is common practice to filter out easy-to-identify abnormal behaviours or system downtime (Section 4.2.2.1) and replace them with missing values (represented as Not Available (NA) values). More advanced methods are developed in dedicated research fields. For instance, [190] provides a fairly complete review of fault detection and diagnosis methods dealing with hot spots, arc faults, short-circuits and inverter failures among other things.

In this section we focus on power loss analysis associated with macro failures (i.e. at the inverter/transformer level), which have a high impact on production rates. The impact of failures at the module or cell levels is negligible in comparison. Some articles reviewed for this work propose methods that detect PV faults from the production time series itself ([191]), while others compare observed production with simulated data (from on-site measurements: [189, 192–194] and off-site observations: [195]) or predicted values ([196]). Such approaches generally assume that defaults are associated with high deviations between measured and simulated values. Most of the statistical approaches are based on normal standard deviation limits ($\pm 1$ SD or $\pm 3$ SD) algorithms ([189, 192, 193]), user-defined statistical thresholds (percentile-based approach developed in [191]) or statistical test analysis comparing the theoretical and the measured power ([194] uses the $t$-test). Regression models are also used; the authors of [196] compare the effective production with estimations derived from decision trees trained with environmental information (i.e. irradiance and temperature). The main issue associated with regression models is the need to possess a sufficiently large training dataset associated with normal operating conditions. Moreover, some approaches are site-dependent: [191] develops a statistical analysis framework based on a statistical clear-sky curve. The methodology identifies the causes of system performance variations (i.e. de-ratings due to shading, clouds and outages) from inverter-based power measurements of rooftop PV systems in semi-arid locations. This kind of climate is well adapted to derive statistical clear-sky curves from previous observations, however it can be difficult to generalise this approach to regions with higher cloud cover rates. To the best of our knowledge, none of the reviewed articles deal with faulty observations in the specific context of PVPF.

Here, we aim to go further into quality control by proposing a robust anomaly detection and correction algorithm even in the presence of noisy data. The option investigated is to identify and potentially correct abnormal production behaviours via proxies of the park production. The first proxies that come to mind are irradiance estimations from satellite-based observations or irradiance modelled with NWPs models. Nonetheless, such sources of information are dismissed inasmuch as inherent modelling errors are present and, in our case, the coarse spatial resolution (and temporal resolution in the case of NWPs) of data prevents a clear distinction between normal and deteriorated production modes. As a result, special attention is paid to on-site observations of irradiance for which we neglect sensors

observational errors but which are subject to measurement defaults. In addition, production measurements at the transformer levels are also considered. The proposed methodology includes two steps; first a preprocess of the inputs makes it possible to reject erroneous observations. Then, a clustering algorithm, requiring a few user-defined parameters, associates each production observation with a production regime.

#### 4.2.2.1 Preliminary control

Some basic cleaning can be performed without any additional information apart from the production time series to identify corrupt data or behaviour deviating from normal operating conditions. This preliminary control is decomposed into three steps:

1. Global checking,
2. Basic quality control,
3. Constant data control.

**4.2.2.1.1 Global checking** Firstly, we check that, (1) Daylight Saving Time (DST) is not implemented in the data, (2) no solar eclipse occurred, and (3) production observations are lower than the installed capacity (given the 15-min temporal resolution of time series, Cloud Enhancement (CE) phenomenon is dismissed).

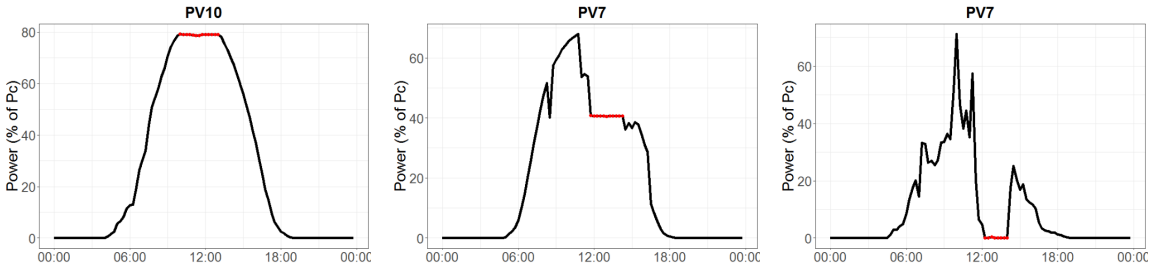
**4.2.2.1.2 Basic quality control** Then, the quality control of the PV power measurements proposed in [189] is applied (the criteria used are summarised in Table 4.1). A low daily mean production may indicate an overcast situation, a potential system failure, or even the covering of the modules by snow. Given the low occurrence of such situations, a visual comparison with production measurements and irradiance forecasts enables us to retain days with high cloud coverage and to flag other days as faulty.

| Label | Criteria | Description |
|---|---|---|
| Upper limit | Comparison of production with extraterrestrial irradiance in the POA | $P_{meas} < 1.5 \cdot I_0 \cdot cos(\theta)^{1.2}$ for $\theta < 85°$ |
| Lower limit | Lower boundary of production measurements | $P_{meas} \geq 0$ |
| Sundown | Production limited to zero during the night | $P_{meas} = 0$ for $\theta^S > 95°$ |
| Daily energy ratio | Days associated with a very low level of produced energy are flagged | $\frac{\sum_{t=1}^{T} P_{meas}}{\sum_{t=1}^{T} P_{sim,cs}} > 0.05$ |

Table 4.1 – Filtering criteria for quality control of the PV production measurements[189].

$P_{meas}$ Measured power normalised by $P_c$ (% of $P_c$),
$P_c$ Installed capacity ($MW$),
$I_0$ Extraterrestrial irradiance normalised to 1 $kW/m^2$,
$\theta$ Incident angle (°),
$\theta^S$ Solar zenith angle (°),
$P_{sim,cs}$ Normalised simulated power in clear-sky conditions (performance and clear-sky models are respectively introduced in Chapter 3 and Section 2.2.1) (% of $P_c$).

**4.2.2.1.3 Constant data control** PV plants can experience total shutdown or curtailment events (i.e. limitation of power output). Curtailments can be mixed up with frozen data phenomena resulting from measurement or recording malfunctions. To avoid assimilating a local shading effect with plant failure, production observations below a 10° elevation angle threshold are disregarded. In both situations, electric production remains constant over a certain period of time. Such behaviours can be easily identified by taking into account the derivative of the production signal. Besides, to avoid misclassifying observation points that slightly deviate from shutdown or curtailment conditions, a degree of flexibility is imposed: $|\frac{dP}{dt}| \leq \epsilon$, where $\epsilon$ is an empirically chosen threshold (here, $\epsilon = 0.02$). Lastly, to be labelled as an anomaly, the constant production must last at least a certain period of time (here, 1 hour is imposed). Figure 4.3 (a)-(b) and Figure 4.3 (c) illustrate two production behaviours identified respectively as curtailment and shutdown.




(a) Curtailment of PV10 plant production.

(b) Curtailment of PV7 plant production.

(c) Shutdown of PV7 plant.

Figure 4.3 – Production anomalies observed on PV7 and PV10 plants. Red points represent production measurements identified as anomalies.

Table 4.2 highlights that the rate of curtailed production is rather low in comparison with null production resulting from shutdowns; PV7 and PV10 are the plants which experience the highest rates of curtailment. Further investigations with the Compagnie Nationale du Rhône (CNR) revealed that for some plants (e.g. PV3 and PV7) curtailment associated with high production rates results from transformer power under-sizing w.r.t. the PV installed capacity. In such a configuration, we assume that the maximum production rate is rarely reached, as a result, production is maximised during morning and afternoon periods but at the risk of a possible saturation at noon. PV9 and PV10 stand out owing to the frequency

of their shutdowns.

| Site Name | Curtailment (%) | | Shutdown (%) | |
|---|---|---|---|---|
| | 2015 | 2016 | 2015 | 2016 |
| PV1 | 0.55 | 0.33 | 0.88 | 0.93 |
| PV2 | 0.00 | 0.08 | 0.55 | 0.63 |
| PV3 | 0.79 | 0.38 | 2.05 | 1.34 |
| PV4 | 0.06 | 0.03 | 1.46 | 0.38 |
| PV5 | 0.33 | 0.27 | 1.02 | 1.15 |
| PV6 | 0.03 | 0.00 | 0.93 | 1.48 |
| PV7 | 1.15 | 0.71 | 0.77 | 0.14 |
| PV8 | 0.71 | 0.38 | 2.79 | 1.34 |
| PV9 | 0.00 | 0.22 | 7.08 | 3.24 |
| PV10 | 1.13 | 0.82 | 4.83 | 14.9 |

Table 4.2 – Duration of estimated anomalies w.r.t. an effective year of production.

#### 4.2.2.2 Additional sources of variability

Grid-connected PV plants are composed of several sub-components responsible for converting irradiance into high-voltage electricity (Figure 4.4). First, PV modules produce Direct Current (DC) electricity from irradiance. As the output voltage of a module is low, modules are gathered into series to form a string. Then, the combiner box brings the output of several strings together into a common bus. The DC output is converted into an AC power through the inverter. Lastly, transformers increase the low-voltage power for grid interconnection purposes. To monitor power production, sensors are set up at several levels of the conversion chain.

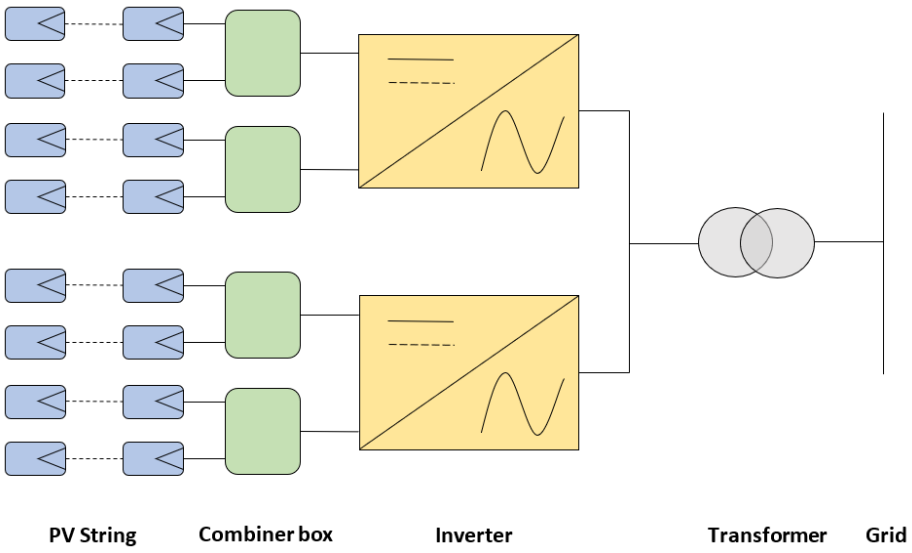



Figure 4.4 – Diagram of a grid-connected PV plant.

In addition to the aforementioned production/measurement defaults, the quality of the

production time series is affected by other factors:

- Missing data in the raw time series resulting from communication or measurement failures at the distribution station level,
- Component shutdowns resulting from scheduled preventive maintenance works,
- Component failures, which necessitate corrective maintenance.

The present dataset is free from missing observations resulting from communication or measurement errors. However, phenomena such as component failure or site downtime, which are independent from local weather conditions, are a source of additional variability and bias which makes it more difficult for the PVPF model to derive relevant forecasting laws. This observation is all the more relevant in an ST context, as a transformer/inverter failure at one site could be viewed as a cloudy situation by a neighbourhood plant. This motivates the identification of behaviours that deviate from nominal operation, and the question of how to deal with them. The easiest approach consists in filtering them out (i.e. replacing the associated values with NA). Depending on the nature of the missing data, this approach, however, can alter the distribution of the dataset and induce bias during the forecast. The other possibility is to implement a corrective approach to approximate the power that would have been produced in nominal operation. This last option preserves the historical dataset (in the sense of the number of observations), which is essential to establish statistical models, but requires additional computational efforts depending on the imputation method used.

#### 4.2.2.3 Issues associated with missing data

Missing data are typically classified into three categories depending on how much they bias the results:

1. **Missing Completely at Random (MCAR)**: concerns all data points for which the probability of being missing is independent from any features in the dataset. The complete dataset (i.e. without any missing points) has the same distribution as the original one. This type of missing point does not introduce bias during model estimation.
2. **Missing at Random (MAR)**: means that missing data points are not related to the missing value itself but do depend on the value of other features.
3. **Missing Not at Random (MNAR)**: refers to data points for which the probability of being missing depends on the value that these points would have taken. In such a situation, the resulting distribution deviates from the original one, which induces bias modelling.

Thus, missing data induce an obvious loss of information but may also degrade forecasting performances by introducing bias in the modelling process if they are not MCAR. As

seen previously, some plants experience a clipping of their power during peak production due to undersized transformers, but in general, plants are more likely to be curtailed close to their rated power due to grid constraint issues. In addition, preventive maintenance works carried out by CNR on power plant parts (i.e. transformers, inverters) are scheduled during wintertime to minimise financial losses. Therefore, such events do not follow MCAR requirements due to their weather dependence, whereas component failures may be associated with a random process.

In [197], the authors study the properties and effects of missing data and imputation methods on the performances of wind power forecasts based on a Vector Auto-Regressive (VAR) model. A preprocessing step rejects data associated with maintenance works and curtailment while individual turbine shutdowns are corrected by re-normalising the site production. The generation of various missing data rates in the training set, mimicking patterns seen in real datasets, highlights that forecasting errors increase at the same time as the rate of missing data. Similar conclusions are reached in [198] for the computation of degradation rates in PV production: as the percentage of MCAR-missing data increases in the incomplete data set, the Absolute Percentage Error (APE) associated with the computation of the degradation rates also increases. Very few studies have applied data imputation to the PVPF field. To name one, [199] proposes an ML framework that enables knowledge transfers from a PV unit to another unit experiencing similar conditions with the aim of filling observation gaps. All of these studies highlight that imputation strategies enhance models' accuracy in comparison with models based on incomplete data sets.

Statistical tools such as mean substitution and imputation based on regression models are usually considered to fill missing data. We opt to employ a physics-based approach considering the proxy data of the production to respect the potential ST dependencies that may exist between the plants.

**Research Gap - *Identification strategy***

The presence of erroneous data tends to bias forecasting models owing to the discrepancy between production anomalies and explanatory features. At this stage, we might wonder whether the bias associated with the presence of erroneous data would be greater than the bias that results from the removal of fallacious data. In such a situation, the identification strategy of spurious observations would be pointless. For the purpose of improving forecast accuracy, is it better to remove or to retain power measured under faulty behaviour?

## 4.3 Identification and imputation strategy

A power plant can operate under three production modes:

1. An *optimal mode* during which all of its sub-components are working properly,

2. A *deteriorated mode* during which some of its sub-components are experiencing failure,
3. A *shutdown mode* when all sub-components are down.

The present section aims at (1) distinguishing the Working Condition Without Failures (WCWF) regime (i.e. optimal mode) from default regimes (i.e. deteriorated and shutdown modes), and (2) refining the partitioning to specify the technical characteristics of each default regime in preparation for correction. To obtain an in-depth quality control of the production time series, the latter alone is no longer sufficient. Thus, we consider a multivariate time series composed of the observed PV production and a variable acting as a proxy of the weather state at the site location. Both features are thereafter denoted respectively as the *target* and the *proxy*. The proxy feature is used to identify deviant behaviours of the target time series. The proposed methodology only considers two inputs (i.e. the time series to analyse and its proxy) but it can be extended to higher dimensions. The approach proposed below possesses several advantages: (1) considered data can be easily retrieved by power producers, (2) it requires low computational efforts, (3) operational forecasts are not compromised because of the possibility to correct data on the fly.

### 4.3.1 Proxies

#### 4.3.1.1 Available data

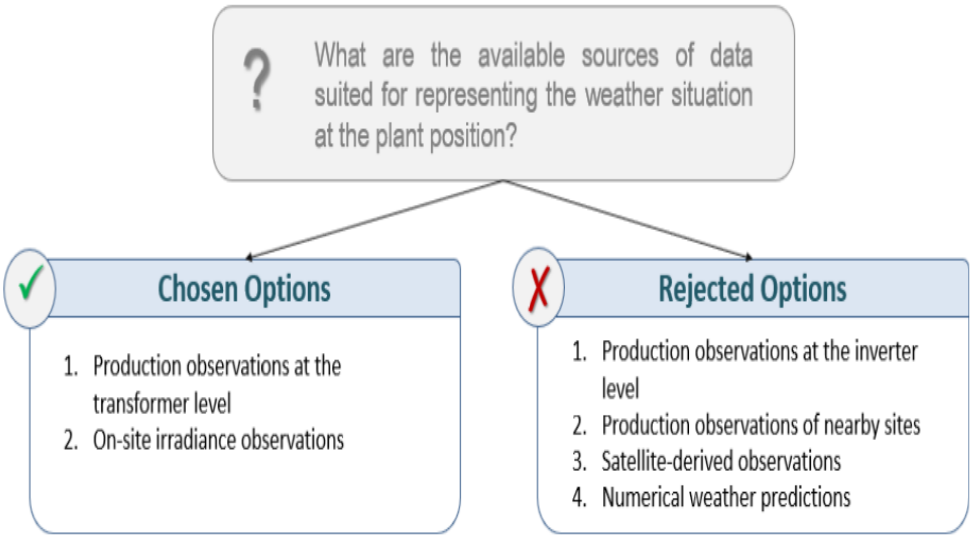



Figure 4.5 – Considered proxies.

Several data sources can be considered to account for the atmospheric situation at the site location (Figure 4.5). A first option could be to turn to Satellite Derived Surface Irradiance (SDSI) and NWPs for their good correlation with the PV production (Figure 2.10). Nevertheless, such data do not allow us to identify abnormal production measurements due to their relative low spatial and/or temporal resolutions. In addition, these sources have been excluded to prevent potential bias during forecasting.

In addition, dependencies between parks have been exploited. Coupled with the investigated model, such inputs are relevant for very close sites but exhibit poor performances otherwise. This conclusion depends on the model used; [199] considers a more robust model based on ST patterns shared by several plants to impute missing data.

As a consequence, we turn to on-site measured parameters. Such a consideration assumes that the spatial variability of irradiance that exists on large-scale installations is overlooked. To obtain a fine vision of the plant's state, production measurements at the inverter level were first considered. Nevertheless, this source was not further investigated due to poor data quality (e.g. non-constant temporal shifts, frozen data, missing values). Ultimately, we consider two proxies: (1) production measured at transformer level, and (2) PV production estimations derived from on-site irradiance observations provided by reference cells and the performance model developed in Chapter 3. The main idea behind transformer-based data is to access the production of each transformer to easily identify deviant behaviour such as transformer shutdowns and even inverter shutdowns. As a result, only sites with at least two transformers are considered (PV3, PV7 and PV9 plants are dismissed).

#### 4.3.1.2 Critical analysis of data

Contrary to electrical power measured at the distribution station, measurements at the transformer stations and outputs from reference cells are not directly exploited by CNR. This leads to a lack of periodic maintenance of sensors and the presence of spurious and missing measurements. As a result, a data integrity control is performed to reject fallacious observations (e.g. frozen data) from both datasets.

The irradiance quality control procedure for Global Horizontal Irradiance (GHI) proposed in [200] is applied with reference cell observations. The key difference between the scope of use of the proposed formulas and our PV installations is that none of them are horizontal. Therefore, the extraterrestrial irradiance in the POA is considered via the use of the angle of incidence. Any reported value that exceeds the limits imposed by Equations 4.1-4.2 is flagged as erroneous and replaced by the NA value (similar mathematical notation to Table 4.1).

- Quality control based on extrema:

$$GTI(t) < min\left(1.2 \cdot I_0, 1.5 \cdot I_0 \cdot cos(\theta(t))^{1.2} + 100\right) \tag{4.1}$$

- Quality control based on rare observations:

$$GTI(t) < 1.2 \cdot I_0 \cdot cos(\theta(t))^{1.2} + 50 \tag{4.2}$$

A visual-based correction of temporal shifts retains the transformer time series of sites PV1, PV5, PV6, PV8, PV10, while the quality of time series of PV2 and PV4 are too poor to be exploitable. The main issue with transformer production datasets is that observations corresponding to a component failure (i.e. $P_{tr} = 0$) or a communication loss

independent from a component failure (i.e. $P_{tr} \neq 0$) are both labelled as null production by the data logged. A simple way to diagnose the origin of these observations is to aggregate all transformer production time series and then compare the resulting value with the power measured at the distribution station. Any deviation from the WCWF regime can be imputed to a communication default and dismissed. This identification is performed by the algorithm proposed in next section.

**Research Gap - *Correction strategy***

In the present work, selected proxies contain erroneous observations that have been rejected. Nevertheless, we cannot guarantee that these datasets are completely free from fallacious data. In this respect, the correction step of the following algorithm will likely generate some uncertainties. As a result, is it better to reject or to correct deviant production observations?

### 4.3.2 Preliminary identification of abnormal regimes

The main idea behind this methodology is based on the strong linear relationship between the target and the proxy features (Figure 4.6). The coefficient of determination R2 indicates the strength of the linear relationship between the observed and simulated power: a value of 1 means that the data perfectly fit a straight regression line. In such a configuration, the regression line characterising the nominal operating mode has a slope very close to 1, which can be easily determined with a Robust Linear Regression (RLR) model. On the other hand, production observations resulting from anomalies are distributed along lines with a lower slope. Potential clusters located above the nominal mode are associated with proxy faulty measurements, and are thus dismissed. Knowing the number and capacity of installed transformers allows us to define production regimes, which can be associated with the different clusters observed in the scatter plot. From there, rejection or correction of data is conceivable.

A clustering approach is implemented to differentiate the various production modes. Such approaches seek to minimise the intra-cluster distance while maximising the inter-cluster distance. First, we draw on [192] which defines a model to identify faults from the system efficiency and in-plane solar irradiance variables. The irradiance feature is binned [2] in such a way that each bin contains the same number of observations. Within these bins, the distributions of system efficiency values are represented by a normal distribution. To define the upper and lower boundaries of the cloud, a 95% confidence interval is used: if the probability of occurrence of an efficiency value is lower than 2.5% (i.e. 1.96 standard deviations below the mean), it represents faulty performance. Lastly, WCWF regime observations can

2. The continuous values of the irradiance levels are placed into "bins", namely ranges of values. In other words, the binning process divides the irradiance feature into distinct groups (e.g. irradiance may belong to the following bins $[0, 100[, \ldots, [900, 1000]$).

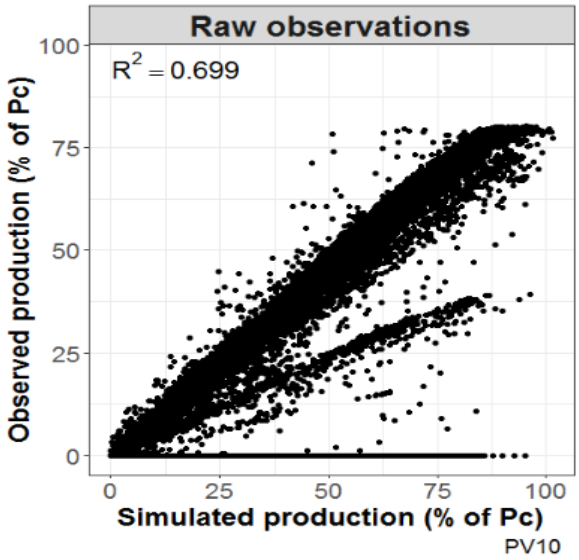



Figure 4.6 – Scatter plot of the normalised observed production w.r.t. the normalised simulated production based on on-site irradiance observations.

be separated from default regime data by combining the results for all bins. The distinctive characteristic of our dataset is that production observations within bins may follow a multimodal normal distribution when default regimes occur. The identification method is then adapted by considering a Gaussian Mixture Model (GMM) to identify the different production modes, based on the assumption that production from normal and anomalous regimes follow different Gaussian distributions. Nevertheless, this approach reaches its limits when it comes to identifying production regimes for low irradiance level (lower left-hand corner of Figure 4.6); in such a region all of the regimes' regression lines are too close to permit a clear demarcation. Similar conclusions were drawn by considering traditional clustering algorithms (e.g. Kmeans algorithm) in other features spaces.

### 4.3.3 Methodology

Thus, clustering the time series into a single unit seems like an attractive option due to its simplicity. However, with such datasets, clustering models perform poorly because the variety of dynamics observed during a year makes it difficult to distinguish the various operating modes. Therefore, the idea is to divide the multivariate time series into smaller segments (typically one or two days) to allow for more detailed clustering. This approach can be assimilated to a sliding window. Then, the temporal segments are transferred into a feature space built from the target and proxy features. This space is composed of the ratio of the target feature over its proxy, and the difference between the clear-sky indexes of the target and proxy features. Within this space, a clustering algorithm gathers observations into groups sharing similar properties. These groups are then associated with production regimes and eventually a correction factor is applied to spurious production observations.

For the reader to better understand the identification and correction processes developed

through this section, a schematic is provided in Figure 4.7. In addition, Figure 4.8 displays a summary of the different options investigated before the end of the proposed methodology.

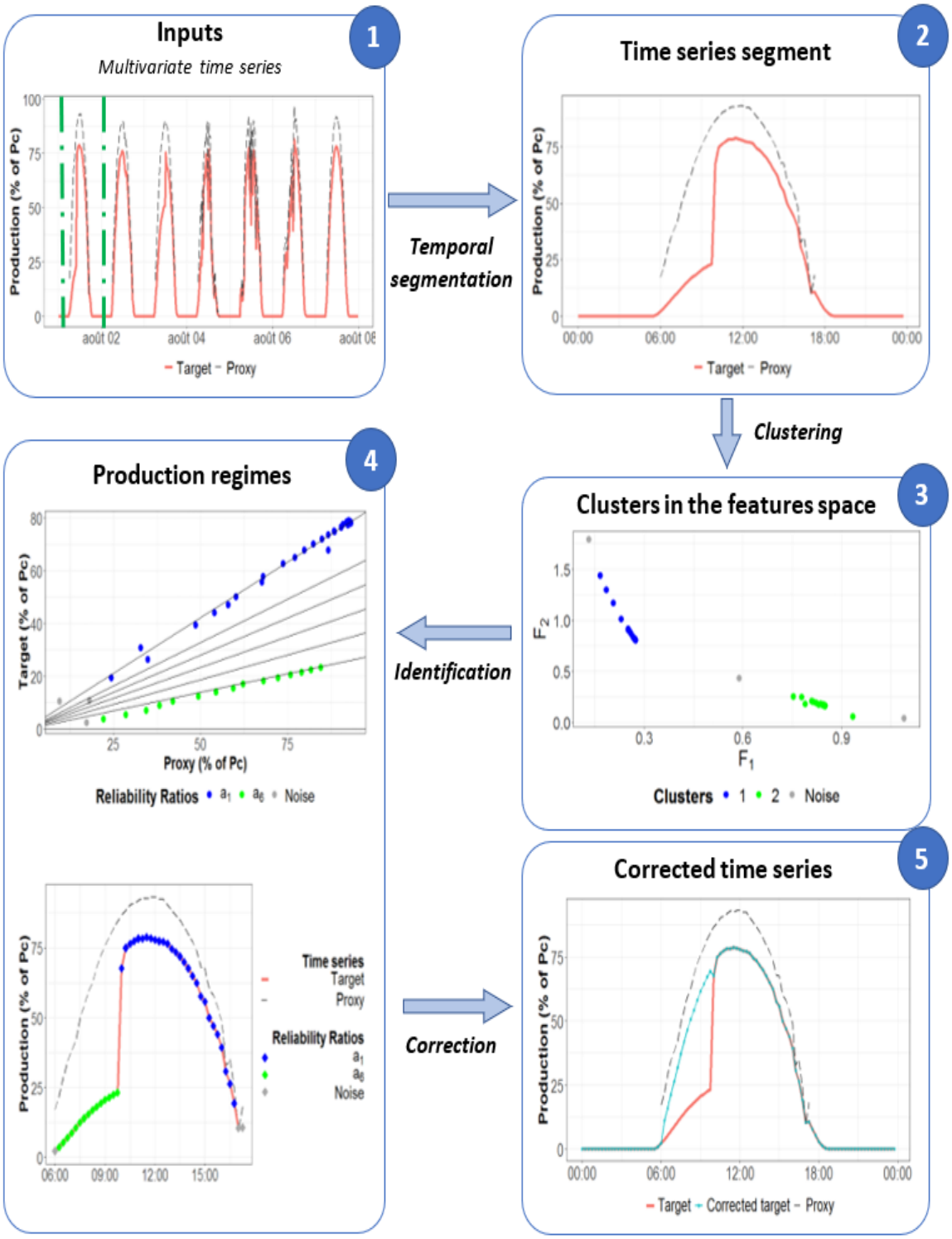



Figure 4.7 – Schematics of the identification and correction processes of data measured during deteriorated or shutdown modes corrected with the reference cell proxy. The identification process of spurious production from transformers observations is slightly different and is detailed in Section 4.3.3.2.2.

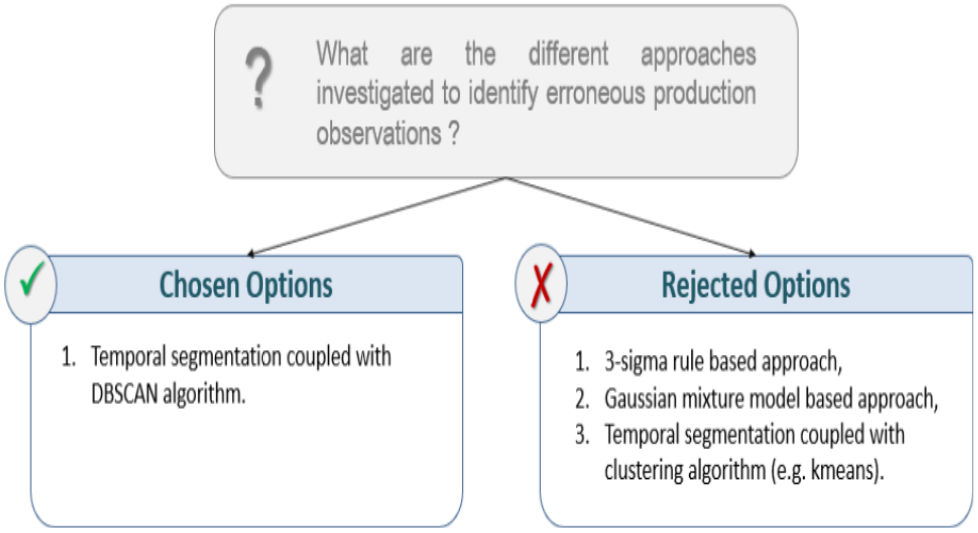



Figure 4.8 – Set of explored options.

#### 4.3.3.1 Clustering algorithm

First, the features space we are working with is constituted by two variables $F_1$ and $F_2$ defined by Equations 4.3-4.4. We observe in Figure 4.7 (3) that $F_2 \in [0, 1.7]$. The highest values reached by $F_2$ are associated with low solar elevation angles. In that case, the irradiance levels measured by the proxy are higher than those simulated by the clear-sky model. This may result from a measurement bias of the reference cell or a poor modelling of the local weather conditions (e.g. atmospheric turbidity) by the clear-sky model. As this behaviour is recurrent and does not impact the clustering approach, no action has been taken.

$$F_1 = \frac{P_{meas}}{P_{proxy}} \qquad (4.3) \qquad\qquad F_2 = \left| \frac{P_{proxy}}{P_{sim,cs}} - \frac{P_{meas}}{P_{sim,cs}} \right| \qquad (4.4)$$

$P_{meas}$ Measured power normalised by $P_c$ (% of $P_c$),
$P_c$ Installed capacity ($MW$),
$P_{proxy}$ Simulated power obtained from the proxy feature (the conversion of irradiance into electric power is developed in Chapter 3), and normalised by $P_c$ (% of $P_c$),
$P_{sim,cs}$ Normalised simulated power in clear-sky conditions (performance and clear-sky models are respectively introduced in Chapter 3 and Section 2.2.1) (% of $P_c$).

In the features space, a clustering method is then applied in order to group observations into meaningful sub-classes according to a certain definition of similarity. For instance, k-means clustering [201] aims at grouping $n$ observations into $k$ clusters by minimising the sum of Euclidean distances between the data points and their respective cluster centroid. Here, we turn to the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm introduced in [202]. This method relies on the notion of density: within each cluster, the density of points is higher than outside the cluster and the density within an area of noise is lower than the density in any of the clusters. The key idea behind DBSCAN

is that for each point of a cluster the neighbourhood of a given radius, $\epsilon$ (Equation 4.5) has to contain at least a minimum number of points denoted as $MinPts$.

$$N_\epsilon(p) = \{q \in D | dist(p,q) \leq \epsilon\} \tag{4.5}$$

| | |
|---|---|
| $N_\epsilon(p)$ | The $\epsilon$-neighbourhood of point $p$, |
| $p, q$ | Two points of the database, |
| $D$ | The database on which is performed the clustering. |

Several classes of points are defined [202]:

- **Core points:** $q$ is a core point if at least $MinPts$ points are within distance $\epsilon$ from it ($Card(N_\epsilon(q)) \geq MinPts$),
- **Directly density-reachable:** $p$ is directly density-reachable from $q$ if ($p \in N_\epsilon(q)$) where $q$ is a core point,
- **Density-reachable:** $p$ is density-reachable from $q$ if there is a chain of points $p_1, ..., p_n$ where $p_1 = q$ and $p_n = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$,
- **Density-connected:** $p$ is density-connected to $q$ if there is a point $o$ such that both $p$ and $q$ are density-reachable from $o$.

A cluster is defined as a set of density-connected points, while outliers are a set of points which do not belong to any cluster. This method does not require any domain knowledge regarding the number of $k$ clusters, and has the ability to identify clusters of any shape as well as outliers. The value of $MinPts$ is a user-defined parameter while $\epsilon$ can be chosen using a k-distance graph [202]. To reduce misclassification rates, observations within a cluster must have a temporal continuity, otherwise they are excluded from the identification and correction process. As a result, parameter $MinPts$ can be understood as the minimal duration of a failure, here we consider $MinPts = 4$ (i.e. 1-hour).

In 2013, a hierarchical clustering extension of DBSCAN was proposed in [203]. Among other things, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is able to consider clusters with different density levels (i.e. different values of $\epsilon$) and requires only $MinPts$ as an input parameter. In the scope of this work, HDBSCAN turned out to be less relevant inasmuch as it tends to identify a higher number of outliers w.r.t. the DBSCAN algorithm. These additional outliers are located at the clusters' border and are identified as such due to a lower density in this area. This misclassification generates correction bias during the correction step. In other words, the PV production observations identified as outliers are not corrected or dismissed during the forecasting approach despite their proximity with default regimes. Due to time constraints, only the feature space developed with the DBSCAN method has been investigated. Conclusions may be different in other features spaces.

#### 4.3.3.2 Regime identification

**4.3.3.2.1 Proxy: reference cells** The previous section clustered the different observations depending on their production dynamics. Now it is necessary to associate each cluster with the corresponding production state of the power plant.

Let a power plant composed of $N_I$ inverters ($N_I > 1$) and $s_i^j$ be the state of operation of the inverter $i$ associated with the operating regime $j$ of the park. Each inverter can either operate in normal conditions ($s_i^j = 1$) or be down ($s_i^j = 0$). Thus, the power plant can experience a total of $N_R = 2^{N_I}$ regimes represented by the set $\mathscr{S} = \{j \in [\![1, N_R]\!], S_j\}$ where $S_j$ is a combination of the different inverters' states. By convention, we impose that $S_1$ and $S_{N_R}$ represent respectively the states where all inverters are on and off. For a specific state $S_j$, the total available capacity, $P_{Site}^{c,S_j}$, is defined by Equation 4.6 where $P_c^{I_i}$ is the capacity of the inverter $I_i$.

$$P_{Site}^{c,S_j} = \sum_{i=1}^{N_I} s_i^j \cdot P_c^{I_i} \tag{4.6}$$

To account for system failures, we introduce the *reliability ratio*, $\hat{a}_j$, which is defined as the ratio of the capacity of the plant experiencing sub-component failures with respect to the plant capacity in WCWF (Equation 4.7). These parameter values range from 0 (i.e. power plant entirely down) to 1 (i.e. absence of failure). With respect to the convention imposed on $S_j$, we have $\hat{a}_1 = 1$, while $\hat{a}_{N_R} = 0$. A plant is composed of several converting units which can have identical or rather close sizing, thus the reliability ratio can take identical or rather close values while they represent different failure configurations. Since knowing which converting unit is on or off does not provide any relevant information, similar values of $\hat{a}_j$ are dismissed except for one. In other words, a selection is performed to reject redundant information.

$$\hat{a}_j = \frac{P_{Site}^{c,S_j}}{P_{Site}^{c,S_1}} \tag{4.7}$$

When the proxy feature accurately represents the PV production process, the parameter $\hat{a}_j$ can be assimilated to the slope of the regression lines of the different production behaviours observed in Figure 4.6. However, in practice the real production regimes slightly deviate from the theoretical ones. This can be explained by processes not explicitly accounted for by the proxy (e.g. conversion losses, simplified modelling). Taking this gap into consideration, a corrective factor is applied to the reliability ratios. This corrective factor is assumed to be independent from the reliability ratios by supposing that losses are proportional to the level of energy produced. Scatter plots comparing the target and the two proxies highlight that the power plants are operating in WCWF for the vast majority of observations. Therefore, the idea is to perform a RLR to derive the apparent reliability ratio of the nominal production mode. In such a context, the corrective factor, $x$, is just

the linear coefficient associated with the proxy variable. The corrected reliability ratios, $a_j$, are then derived with Equation 4.8:

$$a_j = \hat{a}_j \cdot x \tag{4.8}$$

Figures 4.7-2 and 4.7-3 illustrate two production dynamics observed the same day, and associated with two distinct groups during the clustering process described in Section 4.3.3. Outliers typically stand for intermediate values between different production modes. In the next step, the objective is to associate each group of observations with a reliability ratio. To do so, it is necessary to compute the distance, $d(P, a_j)$, between all points $P$ belonging to the same cluster $C_k$, with each reliability ratio $a_j$, $j \in [\![1, N_I]\!]$. The shortest distance between the points, $P$, and the line of slope $a_j$ (namely the perpendicular distance of the points to the line) is computed via Equation 4.9. Then, the average distances, $\bar{D}_k^j$, of all points from clusters $C_k$ w.r.t. each reliability ratio is computed with Equation 4.10. Lastly, each cluster is assigned to the reliability ratio for which it has the shortest averaged distance. In the example of Figures 4.7, *Cluster 1* and *Cluster 2* are respectively associated with reliability ratios $a_1$ and $a_6$.

$$d(P, a_j) = \frac{|y_P - a_i \cdot x_P|}{\sqrt{a_j^2 + 1^2}} \tag{4.9}$$

$$\bar{D}_k^j = \frac{1}{Card(C_k)} \sum_{P \in C_k} d(P, a_j) \tag{4.10}$$

**4.3.3.2.2 Proxy: transformer production** A somewhat different approach is applied when dealing with transformer production observations. In this situation, comparing the proxy with the target offers no information inasmuch as shutdowns are present in both time series. However, the comparison of the production time series of the different transformers provides valuable information regarding the state of production of each transformer, and allows correction of the whole site production through a corrective factor. Power output at the transformer level is not usually used by CNR, as a result, associated sensors are not checked periodically and measurements are subject to malfunction. First, it is necessary to distinguish a power shutdown from a communication loss.

The idea is to use the previous clustering methodology, defined in Section 4.3.3.1, with the target feature (i.e. the production observations of the park) and the aggregation of observations at the transformer level to discriminate shutdowns from communication errors. In the case of effective shutdowns, the production of the site is in line with the aggregated production measured at the transformer level, while gaps are observed for communication losses. In other words, in a 2-D scatter plot built from observed production at the site level and aggregated production at the transformer level, communication losses can be easily identified as the "production regimes" that deviate from the regime associated with the

reliability ratio closest to 1. Thus, communication loss-free observations are located in the cloud of points near the first diagonal.

In a next step, we adopt a more straightforward approach to identify and correct sub-component failures. Subsequent observations that are free from communication losses are then considered in the space composed of features representing the production of the different transformers (Figure 4.9). This representation makes it possible to easily discriminate the various production regimes. The cloud of points along the red line stands for observations associated with a nominal operation of all the transformers, while other line-shaped clouds represent deteriorated production modes. For instance, the cloud along the diagonal of the plane constituted by the $TR01$ and $TR02$ features represents a production mode where transformer $TR03$ is down while the two other transformers are working properly.

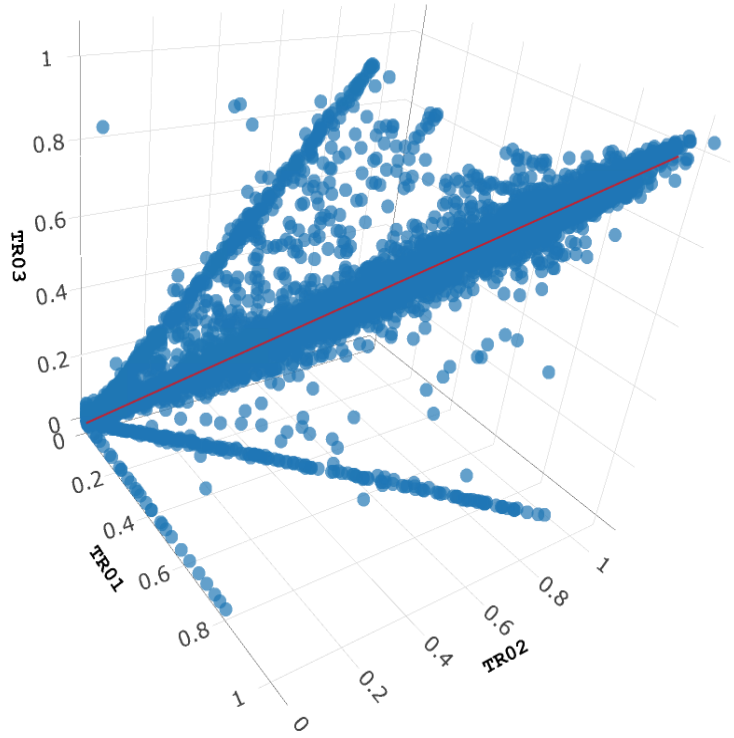



Figure 4.9 – Space composed of the production observations of each transformer at the power plant. The red line represents the nominal mode. Linear groups of points stand for various production modes.

Henceforth, it is necessary to link each group of observations to the relevant production modes. We adopt a regime identification approach similar to what was developed in the previous section dealing with reference cell measurements, except that we do not consider production at the site level but at the transformer level. For a specific state $S_j$, the total available capacity, $P_k^{c,S_j}$, at transformer, $k$, is defined by Equation 4.11.

$$P_k^{c,S_j} = \sum_{l=1}^{N_I^k} s_{k,l}^j \cdot P_c^{I_l} \tag{4.11}$$

$N_I^k$ Number of inverters for transformer $k$,
$s_{k,l}^j$ Operation mode (on or off) of inverter $l$ from transformer $k$ associated with state $j$.

In addition, we introduce the reliability vector defined at Equation 4.12, which gathers reliability ratios at the transformer level and represents the directions of the various abnormal regimes in the transformer features space. To get a better idea of this vector, let us consider Figure 4.9 and assume that the state $j = 2$ represents the situation for which transformer $TR03$ is done. In this specific situation $N_T = 3$, and $a_2^1 = a_2^2 = 1$, while $a_2^3 = 0$. The reliability vector, $\vec{v_2} = (1, 1, 0)$, represents the slope associated with the cloud of points located along the diagonal in the plane $(TR01, TR02)$. The reliability ratio of the whole park for a specific state $j$ is then defined by Equation 4.13.

$$\vec{v_j} = \left(a_j^1, ..., a_j^{N_T}\right), \text{ with, } a_j^k = \frac{P_k^{c,S_j}}{P_k^{c,S_1}} \tag{4.12}$$

$$a_j = \frac{\sum_{k=1}^{N_R} a_j^k \cdot P_k^{c,S_1}}{\sum_{k=1}^{N_R} P_k^{c,S_1}} \tag{4.13}$$

Lastly, a clustering approach attributes points from the transformer features space to the different operating state of the plant, $S_j$. This is done by associating points with the nearest line defined by vector $\vec{v_j}$ (the Euclidean distance is used). For configurations with more than two transformers, the metric distance used is an extension of Equation 4.9 for higher dimensions.

#### 4.3.3.3 Regimes correction

Two proxies have been used to identify abnormal production behaviours. In this section, a methodology to correct the production of the site is proposed.

As each reliability ratio is associated with the number and capacity of operational inverters, it becomes easy to correct observations of the target feature by applying Equation 4.14. This approach can be applied with the production observed at the transformer levels and the reference cells output.

$$P(t) = \frac{P(t)^j}{a_j} \tag{4.14}$$

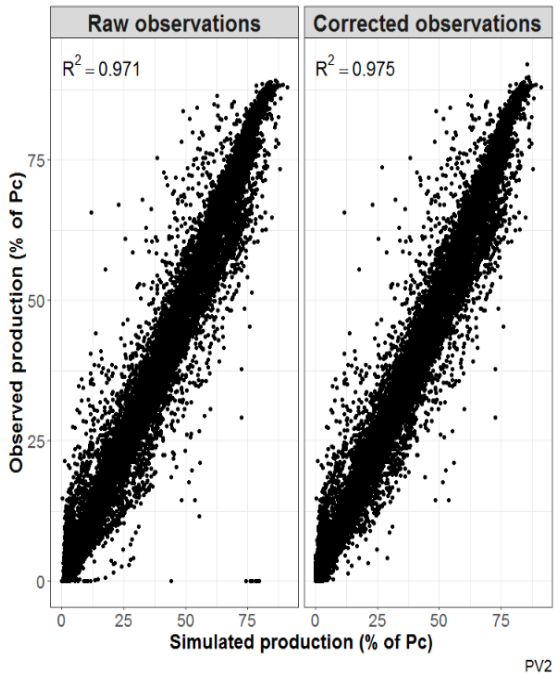
$P(t)^j$ Raw observed production associated with state $S_j$ (where $P(t)^j \neq 0$),
$P(t)$ Corrected observed production.

However, only the reference cells output provides information regarding available resources in case of a complete outage of the power plant. To obtain an estimation of production in this situation, the RLR performed on the dataset is used to convert estimated power from reference cells into the target feature.

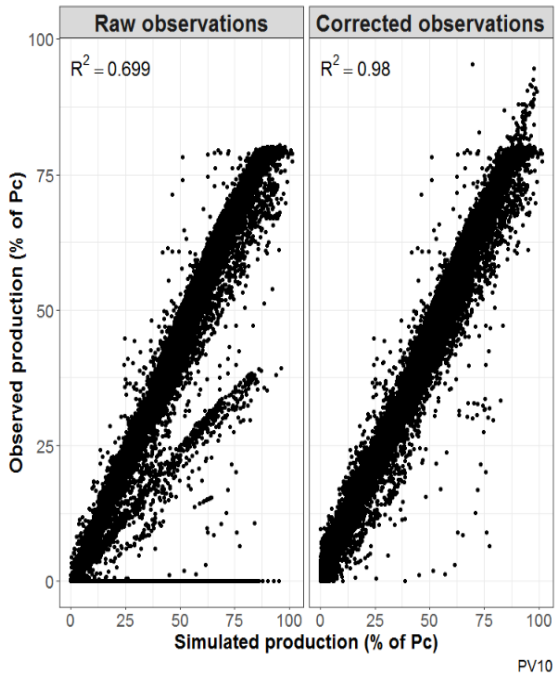
#### 4.3.3.4 Results

**4.3.3.4.1 Example of spurious observations corrected** The aforementioned identification and correction algorithms are successively applied with observations at the transformer level and then with reference cell outputs. This combined approach aims at mitigating the rate of common missing values present in both datasets.

The influence of the correction process is highly site-dependent. Figure 4.10a exhibits very few variations before and after the correction is applied. On the other hand, PV10 (Figure 4.10b), which is more often subject to parts failure, is positively impacted by the correction: observations flagged as abnormal observations (i.e. curtailed observations or production defaults) are realigned with the nominal regime.



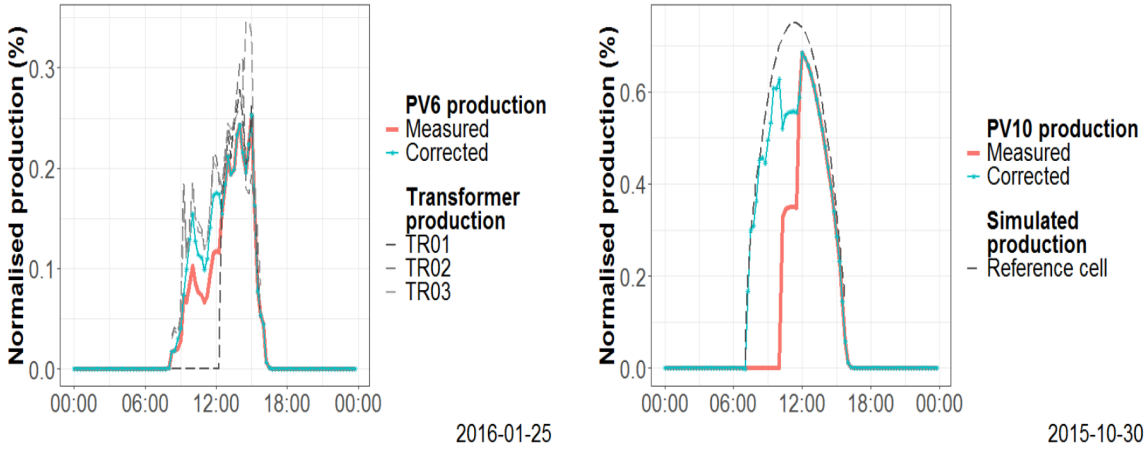

(a) Influence of the correction process over PV2 observations.




(b) Influence of the correction process over PV10 observations.

Figure 4.10 – Evolution of the scatter plots before and after the correction process. The normalised output production and the simulated production from irradiance based on reference cells are represented.

Figure 4.11 highlights the strengths and weaknesses of the proposed methodology and the proxies used. In light of the transformer production observations, Figure 4.11a represents a coherent correction of the production observed at the distribution station. On the other hand, the corrective process is challenged in the case of Figure 4.11b. The on-site measured irradiance allows a fair correction for the beginning of the day (where both transformers are down), then a transformer operates again at around 10h00, but obviously, the production regime of this observation is misclassified and the corrective factor applied is not the correct one.




(a) Curated observations based on the transformer dataset of site PV6.

(b) Curated observations based on the irradiance measurements of site PV10.

Figure 4.11 – Comparison of raw and corrected production observed for a single day.

**4.3.3.4.2 Forecasting performances** To evaluate the impact of the identification and correction processes on forecasting performances, we consider forecasts produced with a forecasting model trained on three distinct datasets:

1. A dataset containing raw observations.
2. A dataset in which observations identified as erroneous are rejected (i.e. replaced by NA).
3. A dataset in which observations identified as erroneous are corrected.

Next, there is the question of the nature of the testing set; should we perform evaluation with a raw dataset or with a dataset in which observations identified as faulty have been rejected or corrected?

We can consider that a raw dataset would be in line with a plant owner's point of view inasmuch as this database represents the actual production of the site, but that it would be to some extent in conflict with a meteorologist's viewpoint because it characterises both the atmospheric state and the technical "reality" of the plant. Here, we assume that production variations that result from technical defaults have to be separated from weather-induced production variations. In an operational context, it is conceivable to reconcile both sources of variability by updating forecasts generated from flawless observations with a corrective factor taking into account the state of production of the plant. Currently, there are no sensors dedicated to measuring the state of production of the different sub-components for the parks under study, but we may assume that such features could be integrated for new projects or as part of retro-fitting measures. Ideally, we aim to generate a dataset that accurately represents the weather situation. However, both the correction and the rejection of observations flagged as spurious have drawbacks. From a statistical point of view, spurious events (e.g. curtailment, maintenance work) do not follow MCAR requirements, and as

such, their rejection modifies the dataset distribution and generates a loss of information. As highlighted in the previous paragraph, the correction process suffers from some flaws. Lastly, we decided to work with a testing set based on corrected observations to preserve the dataset integrity. To some extent, this dataset can be considered as our best representation of the "meteorologist reality".

An Auto-Regressive (AR) model is then fitted on these datasets, while forecasting performances are assessed on a common testing set. Forecasting performances are gathered in Figure 4.12. On the one hand, the raw dataset contains observations associated with WCWF and deteriorated regimes. Deteriorated regimes exhibit lower production levels than the nominal regime. On the other hand, the correcting process applied to the testing set increases the global production levels (because here a correction of production is always associated with an increase in production). Therefore, the level of production observed in the corrected set is higher than that observed in the raw dataset. This explains the normalised Mean Bias Error (nMBE) scores of the models: the *AR(Raw)* model tends to under-estimate production of the testing set compared to the two other models. Both correction and rejection strategies have a positive influence over normalised Root Mean Square Error (nRMSE) scores. Thus, models fitted on the preprocessed observations tend to exhibit fewer large errors than the model fitted on the raw dataset. As the forecast horizon extends, forecasting errors increase. In the case of the *AR(Raw)* model, this tendency is amplified by the model's tendency to under-evaluate production. Conclusions regarding the normalised Mean Absolute Error (nMAE) score are less straightforward. A degradation of the score is observed up to 2 hours lead time compared to the other models. This means that the model based on raw observations tends to make fewer small errors. As the raw dataset is partially composed of faulty production measurements exhibiting a lower variability (e.g. constant level of production for complete shutdowns or curtailed situations) than the two other datasets, we may assume that in the case of the *AR (Raw)* model, more importance is given to the very last observation. In this case, this model behaves like the persistence model, which would explain its good performances for very short-term horizons. We observe similar behaviours with the Random Forest (RF) model (Figure 4.13). In this case, performance variations between the different models are less pronounced but globally the *RF (Raw)* model exhibits the worst scores, except for very short-term horizons but the difference remains slight.

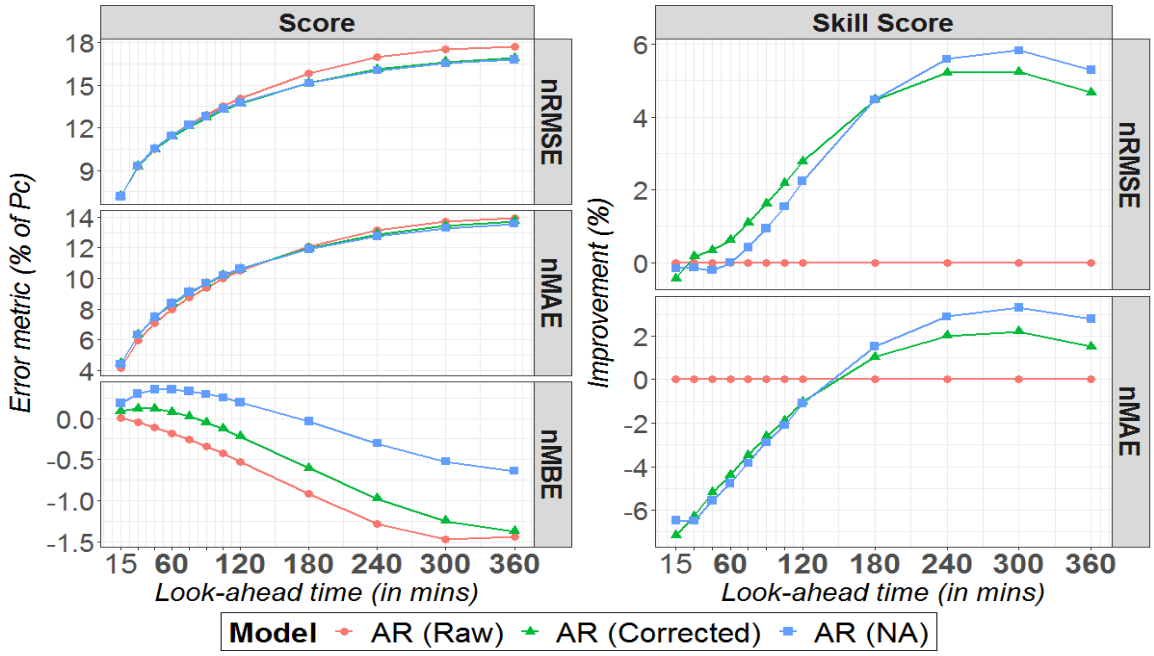



Figure 4.12 – Influence of correction and rejection of abnormal observations over forecasting performances. The AR model is either trained on raw observations or on a dataset whose fallacious observations have been corrected or filtered out.

**Research Answer - *Identification strategy***

For a dataset contaminated with fallacious observations resulting from deviant production behaviours, the rejection of incriminated observations leads to an improvement of the model bias. A significant positive impact is also observed for forecast horizons greater than 2 hours ahead in terms of nRMSE and nMAE. However, a degradation of the nMAE score is observed for shorter lead times when considering the AR model. This ambivalent outcome may be attributed to model-specific features or to the fact that the discrimination process between fallacious and coherent production observations is incomplete. Indeed, the proxy datasets used to identify the different production regimes contain missing data.

Models fed with corrected or filtered-out datasets exhibit rather similar forecasting performances in terms of nRMSE and nMAE, but the latter tends to have a better bias. Despite close scores, Figure 4.14 demonstrates that both forecasts are significantly distinct from each other. Therefore, surprisingly the correction strategy does not provide meaningful additional information compared to the rejection of spurious observations. This may be explained by the fact that the loss of information generated by the missing values is counterbalanced by the noise induced by the correction step.

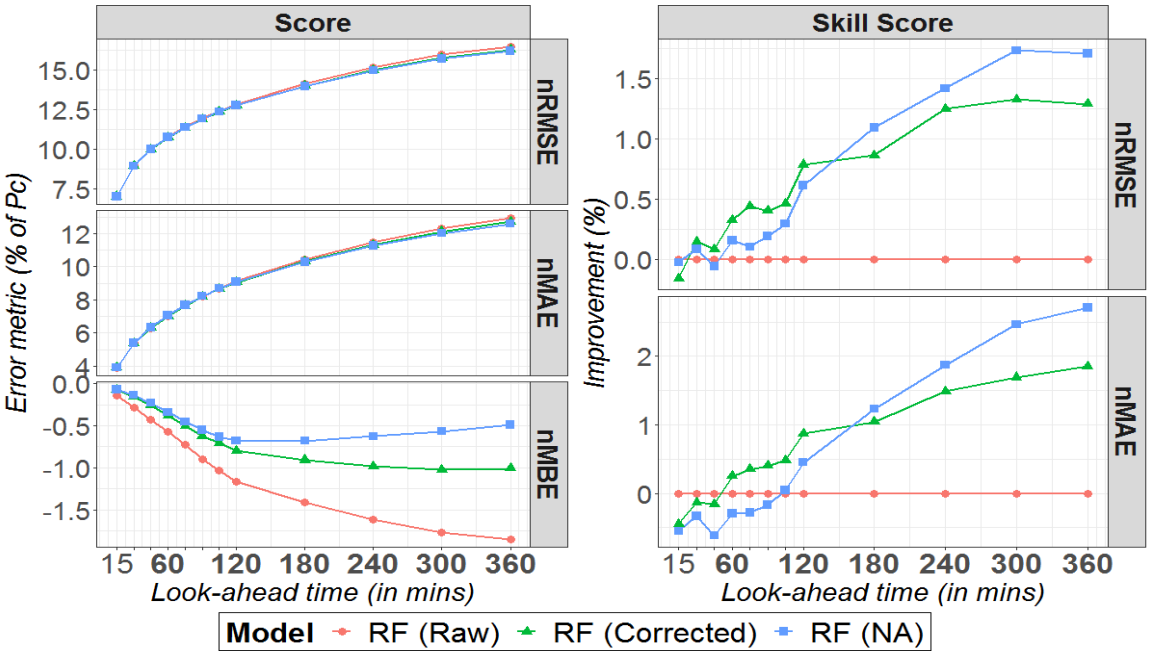



Figure 4.13 – Influence of correction and rejection of abnormal observations over forecasting performances. The RF model is either trained on raw observations or on a dataset whose fallacious observations have been corrected or filtered out.

**Research Answer - *Correction strategy***

In this specific case study, we have spotlighted that the correction or rejection processes of spurious observations associated with abnormal plant behaviours leads to close performances. Due to the alternating performances of the variations in each model, it is difficult to decide which approach is better. As a result, uncertainties generated by the correction process counterbalance the loss of information in the second approach. If plenty of data are available, it is more interesting to reject fallacious observations because the correction process induces extra computation steps.

From now on, the corrected dataset is considered because it contains more observations, which is valuable when employing ST models.

## 4.4 Emphasis of extractable signal information

In the previous section, efforts were made to clean the production signal from the influence of technical issues over its variability. This strategy aims at removing information contained in the signal which is not essential for the generation of forecasts, and so to spotlight the valuable information. In this section we go deeper into the analysis of production time series, and look at solar-related data in a more general way. Solar-related time series are composed of two main components, a deterministic one that represents the Sun's path

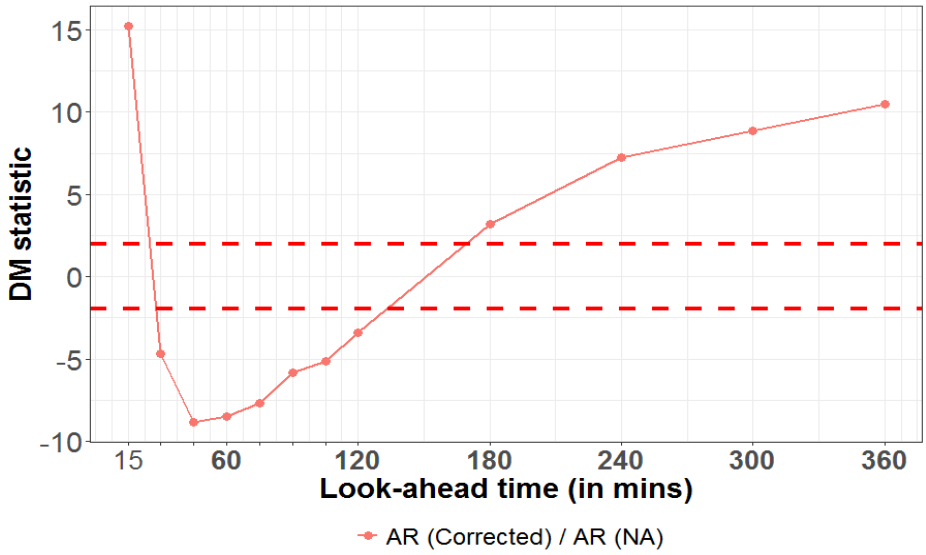



Figure 4.14 – Diebold-Mariano (DM) statistic (defined in Section 2.3.3) between the *AR (Corrected)* and *AR (NA)* models for different forecast horizons. The red dotted lines show the borders delimiting the validation and rejection of the null hypothesis.

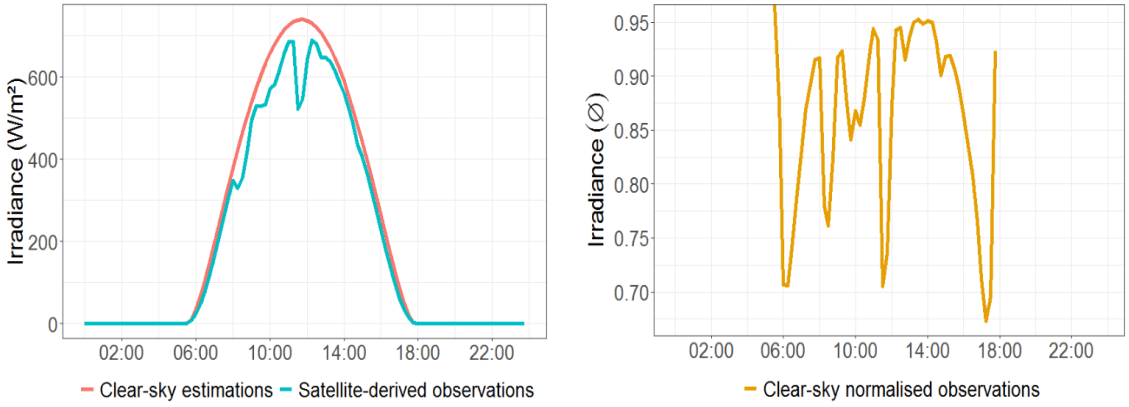
in the sky dome, and a stochastic one, which corresponds to cloud effects. The main idea is to remove the deterministic variability so as to fully focus on the stochastic variability of the time series.

Such an approach also has positive impacts in a context of using ST models. In this paradigm, observations from spatially distributed PV systems are used as a source of additional information that is integrated in the forecasting model. ST models tend to provide forecasts with higher accuracy compared to models based only on endogenous observations. This is made possible thanks to the existence of correlations between the site of interest and the network of Spatially Distributed Units (SDU). In a word, SDU act as a kind of sensor that provides information regarding upcoming changes in weather. Nevertheless, the process of using the power output of one site as a proxy of another is not necessarily straightforward. For instance, first, modules from both sites may have different orientations, which directly impacts the level of received irradiance. Second, sites aligned on the west-east axis tend to have higher degree of correlation due to the Sun's movement in the same direction. Therefore, to get the most from the information contained within spatially distributed data, it is necessary to remove dependencies related to the plant's architecture and environment.

This can be performed with the clear-sky normalisation approach.

### 4.4.1 Clear-sky normalisation

Irradiance that reaches the ground can be viewed as composed of two main variability components: (1) a deterministic one associated with the Sun's motion in the sky dome (i.e. diurnal and annual variation due to the Earth's rotation and orbit), and (2) a stochastic part resulting from weather phenomena such as cloud displacements. A common way to




(a) Observed and simulated clear-sky irradiance.

(b) Normalised irradiance (CSI).

Figure 4.15 – Illustration of the normalisation process.

remove the deterministic component consists in normalising the observed irradiance time series through the concurrent output of a clear-sky model, which estimates the part of solar power reaching the ground assuming a cloudless sky. The resulting feature is called a Clear-Sky Index (CSI) (Equation 4.15). With this methodology, the deterministic part of the production signal is directly modelled by the clear-sky model, while forecasting models are dedicated to the stochastic part.

This index is independent from the Sun's path, as it can be observed in Figure 4.15, which schematically represents the normalisation process. If the measured irradiance and the clear-sky irradiance are both in the Plane-of-Array (POA), then, the normalisation process also removes the dependency on panel orientation (the projection of GHI into Global Tilt Irradiance (GTI) is performed by a projection model - Section 3.3.1.2). Regarding this, in the PVPF related literature, it is common practice to directly normalise the measured production through the ground clear-sky irradiance to avoid the projection and power conversion modelling of GHI. Formally, the CSI is derived from irradiance datasets, but [204] broaden the current definition to PV power by normalising the observed power with the estimated power derived from clear-sky irradiance (Equation 4.16). This approach removes the dependencies on the Sun's path, the system orientation, and the effect of temperature. As a result, $k^{PV}$ parameter acts as a proxy of the cloud cover.

$$k^I(t) = \frac{I(t)}{I^{CS}(t)}. \qquad (4.15) \qquad\qquad k^{PV}(t) = \frac{P(t)}{P^{sim,cs}(t)}. \qquad (4.16)$$

$I$ Irradiance time series ($W/m^2$),
$I^{CS}$ Clear-sky time series ($W/m^2$),
$k^I$ Normalised irradiance time series (CSI) ($\emptyset$),
$k^{PV}$ Normalised production time series (CSI for PV) ($\emptyset$),
$P^{sim,cs}$ Simulated power obtained in clear-sky conditions ($\emptyset$). The model converting clear-sky irradiance into clear-sky "power" is developed in Chapter 3.

#### 4.4.1.1 Clear-sky model

The governing equations of the course of the Sun in the sky dome are perfectly known, but the effects of light crossing the atmosphere are more difficult to model. This has led to the development of numerous of clear-sky models [205]. The simplest approaches model ground irradiance as the extraterrestrial irradiance by assuming a perfectly transparent atmosphere (in such a case, the terminology *clearness index* is consecrated to designate the ratio of radiation measurements to their theoretical values). More advanced approaches are also proposed in the literature. They can be based on a physical modelling of the atmosphere [206, 207] or on a statistical modelling from production observations to bypass the complexity of radiative transfer [189, 191, 208]. The latter option can be appealing in some aspects inasmuch as it does not require the implementation of the decomposition/projection, and power conversion modelling steps.

In [107], the authors provide a fairly complete comparison of some of the most highly cited clear-sky solar radiation models. This review is based on 36 validation studies and several versions of around ten clear-sky models. Due to methodological difficulties, the authors were not in a position to draw general guidelines regarding best practices according to the climatic conditions of the area of interest. However, they highlight that the use of the Linke Turbidity (LT) factor [3] in some clear-sky models (such as the European Solar Radiation Atlas (ESRA) [207], which is based on climatological monthly means of LT) is a significant source of uncertainty. Similar conclusions are reached in [205]: the simplicity of LT-based models does not compensate their limited accuracy and lack of universality. In addition, this article represents a comprehensive validation study of 75 clear-sky irradiance models tested with worldwide irradiance measurements from 75 ground stations. For temperate climates, the MAC2 [210] and REST2v9.1 [211] models provide the best performances. MAC2 is the best overall for all of the climate regions considered, despite being one of the simplest models evaluated. Another advanced clear-sky model, McClear [212], developed within the Monitoring Atmosphere Composition and Climate (MACC) project, is also represented. In brief, this model aims at accounting for the optical state of the atmosphere intraday variabilities by integrating concentration observations of some atmospheric components (e.g. the total column ozone or the total precipitable water vapour). McClear ranked 35$^{th}$ globally in [205].

Similar findings are reached concerning the clear-sky model's complexity for solar forecasting applications in [213]. This study compares the forecasting performances obtained with a naive reference method fed with CSI issued from three clear-sky models namely; the average (Ineichen–Perez-[214]), good (McClear-[212]) and best (REST2-[211]) models according to the classification derived in [205]. The study highlights that there is no evidence proving that high-performance clear-sky models are superior to simpler ones in the forecast-

3. The LT factor approximates the atmospheric absorption and scattering of the solar radiation under clear skies [209].

ing domain. However, [215] concludes that the McClear model is superior by comparing the latter with the ESRA model to estimate solar irradiance from satellite observations, before using a Cloud Motion Vector (CMV) model to derive forecasts.

In the framework of this thesis, we have at our disposal the McClear and ESRA models. The direct implementation of clear-sky models is not trivial, especially because a lot of inputs need to be measured (i.e. total column ozone, aerosol optical depth) [216], and thus the easy access to McClear output is welcome (outputs are accessible via a free online service [4] that simply requires registration). Other advanced models such as MAC2 and REST2 have been dismissed due to the need to access meteorological parameters. A preliminary comparison between both models is performed in Appendix B, where an AR model is fitted with production time series normalised by the two clear-sky models. Findings show that the AR model fed with normalised production exhibits a lower nRMSE when coupled with the McClear model. Subsequently, clear-sky time series derived from the McClear model are considered in this study.

#### 4.4.1.2 Clear-sky index

The CSI and clear-sky models are widely used in the solar-related literature (e.g. to compute the components of ground irradiance, to derive irradiance from satellite observations) and in the forecasting domain closely linked with time series. Regarding this, clear-sky models allow the derivation of a naive forecasting approach based on the persistence of the CSI (this point is detailed in Section 2.3.1). The values of the CSI also provide qualitative and quantitative information regarding the cloud types and their distribution: values close to 1 indicate clear-sky conditions, while lower values are associated with different degrees of overcast situations. Such an analysis is performed in Figure 2.6 from on-site irradiance observations. The CSI typically follows a bimodal distribution (Figure 4.16) with contributions from cloudless situations (i.e. the main mode is located near $k^{PV} = 0.9$), and cloudy situations (the second mode is achieved at $k^{PV} = 0.2$). This pattern is observed and modelled in the literature ([217, 218]). Overall, the CSI derived from all of the sites under study follows the same pattern. The CSI also offers the possibility of identifying shading events [204]. However, such an approach has not been investigated due to a normalisation issue: aberrant normalised quantities (i.e. $k$ or $k^{PV} >> 1$) are observed for low solar elevation angles.

We observe in Figure 4.17 that low solar elevation angles lead to artificially large $k^{PV}$ values (similar behaviour is observed with $k^I$). This may result from (1) numerical instabilities of the CSI computation, (2) CE, (3) a bad modelling of the clear-sky, or (4) a lack of accurate information regarding the plant's geometry. The CE phenomenon is detailed in Section 3.3.1.1. In brief, during a cloudy period, CSI can be greater than 1 due to light

4. http://www.soda-pro.com/. At the time of writing, this website provides us with past estimations up to current day-2. In an operational context, it is possible to generate forecasts of clear-sky irradiance from a personal implementation of the McClear model fed with forecasts of relevant quantities.

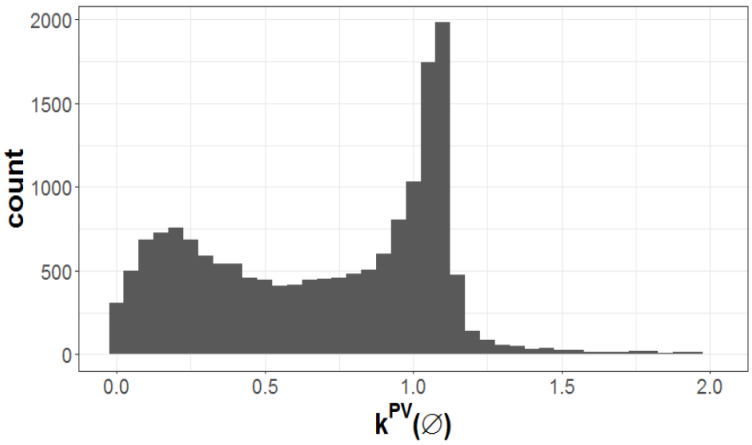



Figure 4.16 – Histogram of the CSI for PV and its bi-modal distribution (PV7). To facilitate visualisation, high values of $k^{PV}$ are truncated.

reflections from cloud edges. In these conditions the power of PV modules can be 30% higher than in standard test conditions [189]. In the present situation, the temporal amplitude of this phenomenon is not large enough to explain the high $k^{PV}$ values (typically CE events last 20-140s [189]). The other reason often invoked in the literature is the poor clear-sky modelling for early morning and late afternoon hours [115, 219]. This flaw could be explained by the fact that clear-sky models mainly consider integrated irradiance over the whole spectrum; still, [220] highlights that the greatest spectral differences occur at dawn/dusk or under heavy cloud cover. Besides, PV response depends on the light wavelength. In addition, under-estimations may also result from an approximate knowledge of the module's orientation angles, which biases the projection onto the POA, thereby increasing the uncertainties during the normalisation process.

The great majority of studies dealing with CSI prefer to reject data associated with low elevation angles [61, 204, 219] (typically a 5° threshold is chosen) on the grounds of limited energy produced during this period. Other approaches, such as [58], try to overcome the normalisation weaknesses for low solar irradiation by proposing a statistical correction, which is also beneficial for stationary properties. The ambition to provide operational forecasts with a production schedule even for early and late hours of the day motivates [115] to keep low irradiance data. Contrary to [115], which considers performance models, here, we assume that high CSI values may bias forecasts performed with low-robust algorithms. Thus, within the scope of this thesis, we decided to reject values observed for solar elevation angles lower than 5°.

### 4.4.2 Signals dependencies

This section aims at assessing the dependencies present in the production observations of a power unit as well as dependencies that may exist among distributed plants. Here, the term *dependence* refers to the fact that the production of two sites may be correlated in the

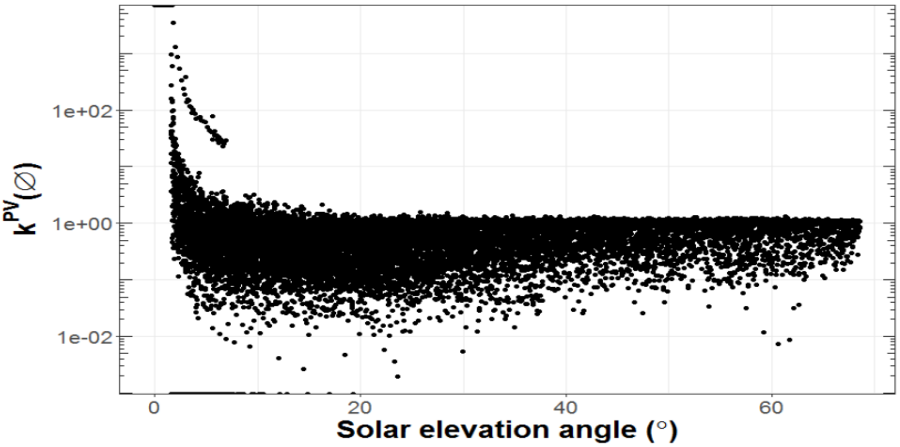



Figure 4.17 – CSI for PV as a function of the solar elevation angle. The sequence of null values is assumed to result from shading effects or excessively low irradiance levels to engage the inverters.

sense that they both experience similar weather patterns with some temporal delay. Thus, to assess such dependencies, it is necessary to quantify the degree of similarity between PV outputs' underlying patterns. To do so, several approaches are conceivable: we can think of distance measures such as Euclidean distance or dynamic time warping [221]. Here, we choose to focus on the correlation coefficient.

Real data exhibit two main properties: correlations (i.e. presence of a dependence structure between two variables or between a signal and a delayed version of itself), and periodicity (i.e. the repetition of a certain pattern at regular time intervals). Auto-Correlation Function (ACF) and Cross-Correlation Function (CCF) are widespread tools that provide an insight into the correlation dynamics of data. However, these approaches were designed to identify correlations in stationary and linear data. When applied with out-of-scope data, such tools can be misleading. Let us assume a time series, $X_t$, with a large upward trend and a high value of $x_t$ is likely to be followed by a high value of $x_{t+1}$, which results in a high auto-correlation. The latter is mainly due to the non-stationarity rather than the actual auto-correlation. As a result, detrending becomes vital to properly analyse correlations by avoiding the manifestation of fake correlations and by magnifying genuine dependencies [222]. Such approaches are referred to as Detrended Cross-Correlation Analysis (DCCA).

The presence of the deterministic component associated with the Sun's path increases the cross-correlation between irradiance-based signals (and inter-correlation of the production signal) by diluting the impact of stochastic weather variations. Such characteristics are thought to negatively impact the derivation of statistical laws and feature selection.

#### 4.4.2.1 Stationarity

Irradiance-related features (e.g. irradiance, PV production) are non-stationary by nature, which makes them more complex to investigate while reducing the set of statistical tools available in time series analysis [46].

In simple terms, stationarity means that the statistical properties of a time series do not change over time. Different degrees of stationarity are defined in the literature; a time series is said to be strictly stationary if the joint probability distribution function, $F$, of the stochastic process, $\{X_t\}$ is invariant under translation (Equation 4.17), while weak stationary implies that the mean and the autocovariance do not depend on time and the $2^{\text{nd}}$ moment is finite.

$$F(x_1, ..., x_t) = F(x_{1+h}, ..., x_{t+h}) \forall h \tag{4.17}$$

This property is often required by time-series-based forecasting models, like those from the Auto Regressive Moving Average (ARMA) family. One generic approach to make time series stationary consists in differentiating it (i.e. computing the differences between consecutive observations). Such a trick is part of the Auto Regressive Integrated Moving Average (ARIMA) modelling strategy. Nevertheless, not all time series can be differentiated to achieve stationarity because some may exhibit a strong seasonal behaviour such that the entire auto-correlation structure of the series depends on the season [223]. This point is verified experimentally by [58], who highlights that differentiation is not sufficient to remove non-stationarities from irradiance-related features. Several approaches are proposed to deal with the non-stationary behaviour of PV production time series. The seasonal decomposition procedure proposed in [224] breaks down a time series into seasonal, trend and irregular components. After the decomposition, the seasonal cycle and the trend are subtracted from the irradiance time series. Such an approach is used to feed an ARIMA model in [225]. One can also find approaches based on Wavelet Transform (WT), which is a signal pre-processing tool that decomposes PV production at different timescales [226–228]. For further information regarding wavelet decomposition the reader may refer to [229].

It is common practice in the PVPF-related literature to resort to stationarity preprocessing tools when dealing with ARMA-based models. On the contrary, such approaches are not widespread in the field of ML-based forecasting (e.g. RF, Artificial Neural Networks (ANN)). To the authors' knowledge, only a few studies broach the subject: [46] assumes that such a data preprocessing step might be beneficial for ANN by providing trend-free time series; in [230], the clear-sky normalisation process is considered but mainly for its normalisation property. In [213] the author remarks with humour that time series forecasters and ML users have still not reached an agreement over the utility of normalising. The latter faction assumes that a well-trained ML algorithm is able to automatically determine the deterministic trend of the time series (i.e. its seasonal components) despite the Occam's razor.

**Research Gaps - *Normalisation process/ML***

Thus, no clear evidence of the influence of data stationary properties over ML-based forecasts has been found in the literature. Can the normalisation process be valuable for forecasting performances?

#### 4.4.2.2 Evaluation of the normalisation process

In [58], the authors assume that the inclination does not affect the stationarity process, since a corrective factor $\eta$ converts clear-sky GHI into simulated power. Formally, the resulting CSI is deprived from its seasonal dependence, while systems dependencies (i.e. inclination, temperature) remain. Moreover, the projection of GHI onto the POA is a much too intricate process to be simplified linearly (this process is detailed in Section 3.3.1.2).

**Research Gap - *Normalisation process/CS model***

We already know that even the best clear-sky models are not able to produce a stationary CSI time series [213], but we might wonder whether it would be possible to improve the degree of stationarity of the CSI by considering clear-sky series integrating the plant's geometry as well as local weather conditions (e.g. temperature)?

To answer this question, we define three CSI:

$$k^1(t) = \frac{PV(t)}{GHI(t)}, \quad k^2(t) = \frac{PV(t)}{GTI(t)}, \quad k^3(t) = \frac{PV(t)}{PV^{sim,cs}(t)} \tag{4.18}$$

**4.4.2.2.1 Qualitative inspection** First, a visual inspection is performed to identify periodicity, trend or change in variance. Figure 4.18 gathers the three aforementioned CSI. The figure displays a yearly periodicity pattern for $k^1$, while nothing apparent is observed for the other two CSI, apart from a few sparks. However, it seems that the different CSI all exhibit trends and changes in variance due to seasonal variations of weather states. Indeed, in general, weather states observed during autumn and winter are mainly characterised by overcast situations with a significant intraday variability, while summertime weather states are mainly composed of clear-sky days with low intraday variability.

**4.4.2.2.2 Statistical test** Next, we turn to statistical hypotheses to test for stationarity in the CSI time series. Stationarity tests allow us to check whether a series is stationary or not. In the literature, popular tests include Kwiatkowski–Phillips–Schmidt–Shin (KPSS) [231], and Augmented Dickey-Fuller (ADF) [232], which are unit root-based tests [5]. Here we focus on the KPSS test, for which the null hypothesis ($H_0$) stipulates that the series is trend stationary (i.e. the mean can be growing or decreasing over time), while ($H_1$): the series has a unit root (i.e. it is non-stationary). From Table 4.3 we observe that for the three CSI considered, the null hypotheses of the KPSS can be rejected. As a result, the KPSS test reports that the series are not stationary. Further investigations highlight that differentiating the CSI leads to stationary conclusions from both the KPSS and ADF [6] tests,

5. Let $\Phi(B)$, the polynomial notation of the backshift operator $B$ (i.e. this operator shifts the data back in time: $B(y_t) = y_{t-1}$), the process $Y_t$ is not stationarity if the $\Phi(B)Y_t$ polynomial has a root equal to unity.

6. The hypotheses for the ADF test are ($H_0$): the series has a unit root (i.e. the time series is non-stationary), and ($H_1$): the time series has no unit root and so the process is stationary.

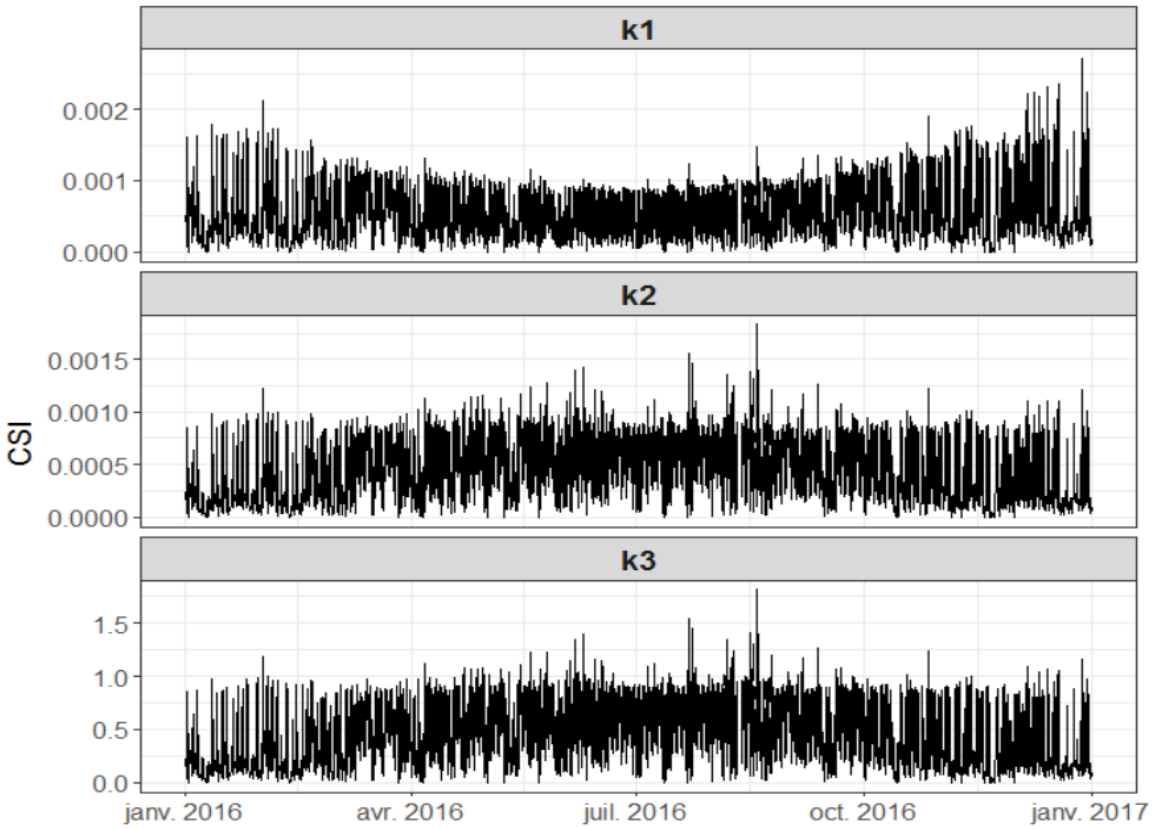



Figure 4.18 – Time series of the CSI considering three normalising strategies (Equation 4.18). PV7 is considered.

which indicates that these series are difference-stationary. These conclusions encourage the use of integrated time series (i.e. $\Delta y_t = y_t - y_{t-1}$), but require the computation of the $h$ intermediate increment forecasts ($\widehat{\Delta y_{t+1}} \dots \widehat{\Delta y_{t+h}}$) to derive $\hat{y}_{t+h}$.

| CSI | KPSS (0.146) |
|---|---|
| $k^1$ | 2.73 |
| $k^2$ | 1.61 |
| $k^3$ | 1.54 |

Table 4.3 – Stationary tests studied at the 5% confidence level. For the KPSS test, the null hypothesis ($H_0$) stipulates that the series is trend stationary, while ($H_1$): the series has a unit root (i.e. it is non-stationary). If the value of the test result is greater than the critical value (in parentheses), the null hypothesis can be rejected in favour of the alternative hypothesis at the 5% level of significance, otherwise, it is accepted.

This step by step approach can be time consuming inasmuch as we consider 15-min time-step data and forecast horizons up to 6 hours ahead. This motivates us to avoid time series differentiation and to further explore the stationary properties of CSI, and especially local stationary time series[7]. To do so, we follow the methodology proposed in [213]. The main idea consists in comparing two samples to determine whether they follow the same distribution. For the same CSI, different conditional distributions are generated and compared pairwise via the two-sample Kolmogorov–Smirnov (KS) test. This non-parametric

7. Non-stationary time series with statistical properties that change slowly over time.

test tests whether two samples come from the same distribution. Figure 4.19 reveals that most of the KS tests reject the null hypothesis in the case of the $k_1$ CSI (i.e. most of the conditional distributions are not identical). On the other hand, $k_2$ and $k_3$ exhibit slightly higher proportions of adjacent bins for which the null hypothesis is retained. This highlights the slow evolution of the statistical properties and may presume local stationarity. The property of local stationarity is a key assumption of Time Varying AutoRegressive (TVAR) processes [233]. Chapter 6 proposes a way to derive time-dependent models based on a physical characterisation of the atmosphere.

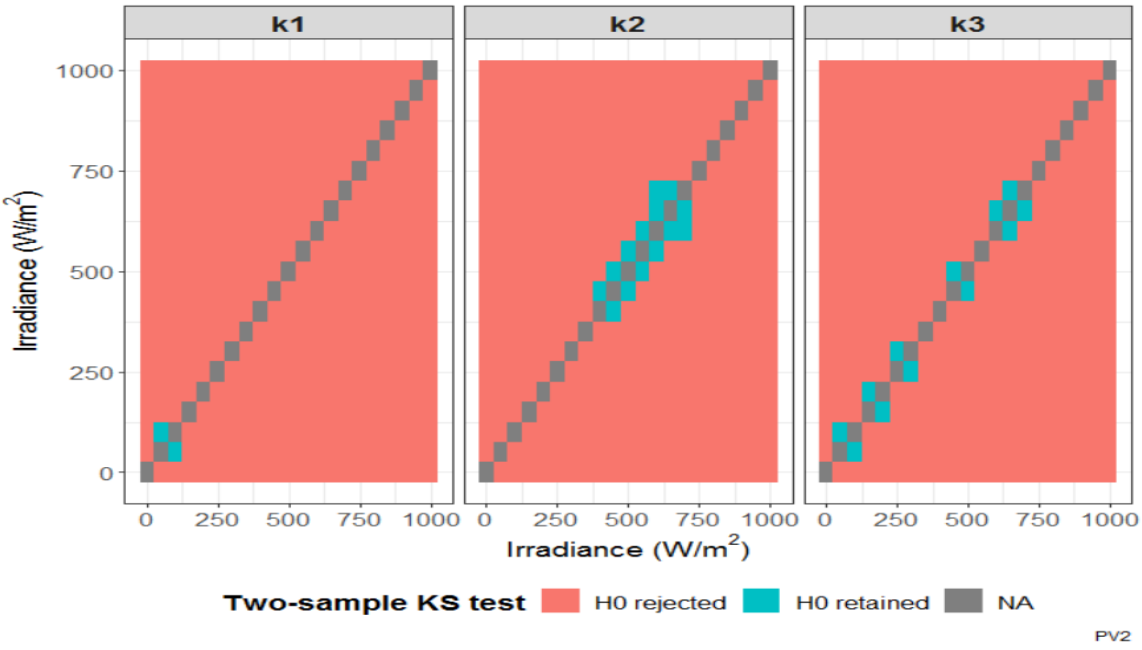



Figure 4.19 – Output of the two-sample KS test for pairwise comparison of conditional distributions of CSI given the level of irradiance at 5% level of significance. ($H_0$): the samples are drawn from the same distribution.

**4.4.2.2.3 Impact over forecasting performances** Previous analyses provide valuable information regarding the statistical properties of CSI time series. To discriminate the three normalisation processes associated with the CSI ($k^1$, $k^2$, $k^3$), forecasting performances are computed.

Figure 4.20 represents forecasting performances obtained with the AR model. It is obvious that the forecasting model based on the $k^3$ CSI outperforms other approaches in terms of nMAE. Conclusions are less straightforward considering the nRMSE score: forecasting performances of the three approaches are rather close; notwithstanding, the model based on the $k^2$ CSI slightly outperforms its counterparts.

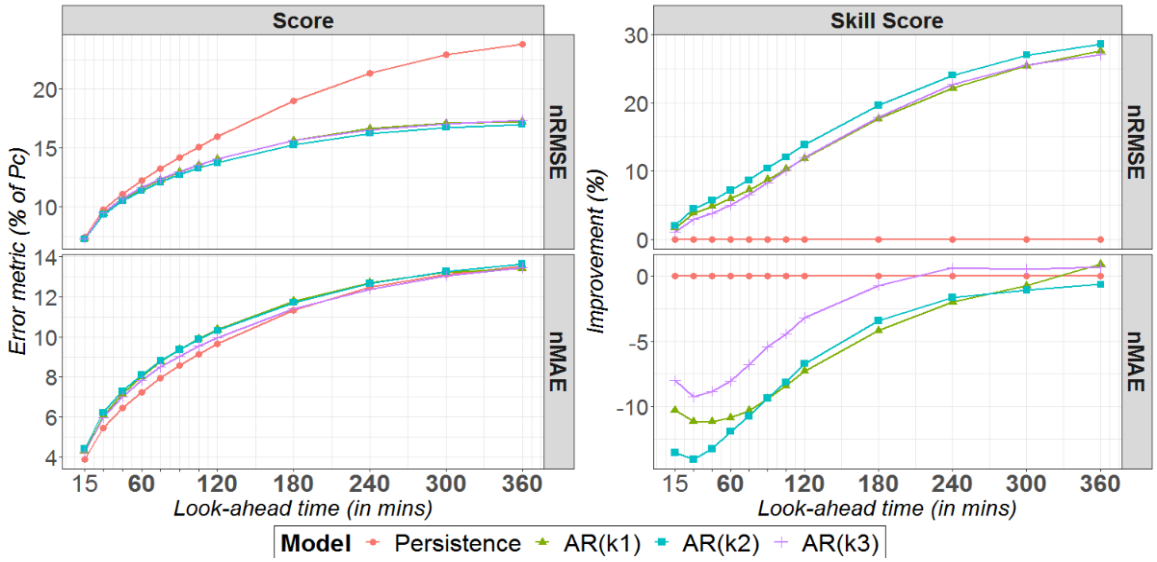



Figure 4.20 – Forecasting performances derived with the AR model considering three CSI, namely $k^1$, $k^2$, and $k^3$.

**Research Answer - *Normalisation process/CS model***

The three alternatives proposed to normalise production turned out to be unable to provide stationary time series. However, the normalisation approaches based on the irradiance-projection and performance models seem to exhibit local stationarity properties. Further investigations spotlight that the AR model based on production observations normalised either with the clear-sky irradiance projected on the POA or with the simulated clear-sky power exhibits performance improvements compared to a model whose inputs have been normalised by the clear-sky GHI.

Subsequently, the normalisation process based on the clear-sky simulated power is retained.

Henceforth, we investigate the impact of the clear-sky normalisation approach on ML techniques. Forecasting models are either fed with raw production observations or with normalised clear-sky inputs. On the one hand, the latter approach explicitly removes the Sun's movement patterns from the production signal before the regression process, but reintegrates it later during the de-normalisation process (Figure 1.10). Thus, this kind of model focuses on cloud-induced variability. On the other hand, the raw production observation-based approach has to infer the Sun's patterns as well as the impact of clouds on production.

On the whole, Figure 4.21 highlights that the clear-sky-based normalisation process clearly has a beneficial impact on the forecasting performances of the RF model. Indeed, the *RF(k3)* model outperforms the *RF(PV)* model, both in terms of nRMSE and nMAE,

whatever the forecast horizons under study. However, the large difference in accuracy between the *RF(PV)* and *RF(k3)* models is quite surprising.

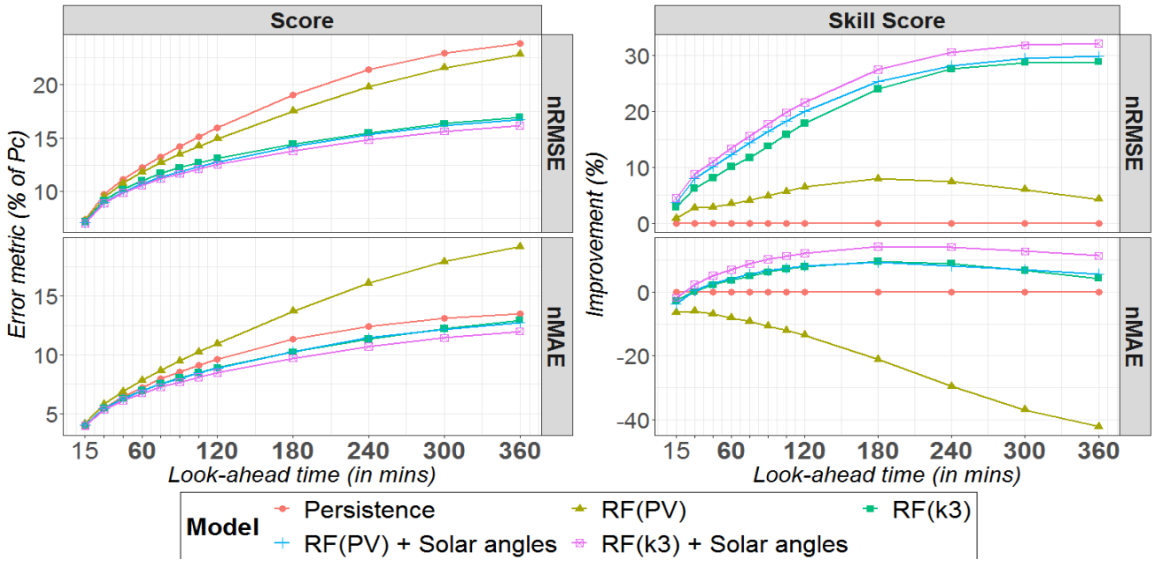



Figure 4.21 – Forecasting performances of an RF model fed with clear-sky normalised data (i.e. *RF(k3)*, *RF(k3) + Solar angles*) or non-normalised inputs (i.e. *RF(PV)*, *RF(PV) + Solar angles*).

At this point, it is necessary to point out that performances are assessed with forecasts generated from sunrise to sunset. Yet, depending on the lead time considered, forecasts for the early morning may be generated with nighttime production observations. In this context, early morning forecasts may be quite inaccurate due to the lack of recent diurnal observations. Thus, it is relevant to distinguish both types of forecast: Figure 4.22 represents forecasting skill scores obtained with forecasts generated either with nighttime or daytime observations. In other words, the left graph represents the accuracy of models fitted with observations before sunrise, while the right graph is obtained with models fed with, at least, the first observations after sunrise. We observe that the main gap between the *RF(PV)* and *RF(k3)* model performances results from nighttime-generated forecasts. In this context, the *RF(k3)* model only has to deal with signal variations due to clouds, in so much as it tends to predict observed CSI trends, learnt during the training step (Figure 5.23). Then the de-normalisation process allows us to explicitly "add the solar curve to the forecasts". In the case of the *RF(PV)* model, this model has to infer simultaneously the influence of clouds and the Sun's path on the production profile. Given the low skill scores, the model struggles to do so. When daytime observations become available, the accuracy of the *RF(PV)* model remains low despite the solar profile being embedded in the production profile. This may be explained by (1) a too restricted training set (only one year of observations is used, which may be not sufficient to derive statistical laws regarding the Sun's path), (2) the model is not complex enough to deal with the different variability patterns, or (3) a lack of informative inputs or easy extractable information.

To help the model perform better, we add features that explicitly characterise the posi-

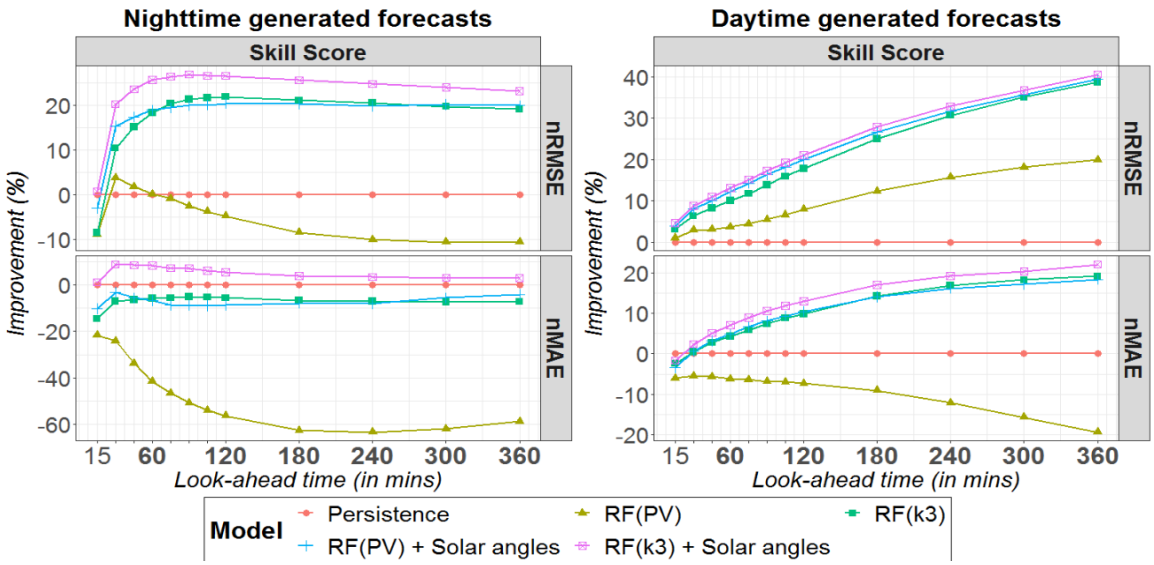



Figure 4.22 – Skill scores, w.r.t. the persistence model, of an RF model fed with clear-sky normalised (i.e. *RF(k3)*, *RF(k3) + Solar angles*) or non-normalised inputs (i.e. *RF(PV)*, *RF(PV) + Solar angles*).

tion of the Sun (i.e. elevation and azimuth angles) at the target time (i.e. at time $t + h$). The inclusion of this information improves forecast accuracy both for nighttime and daytime generated forecasts to such an extent that the resulting performances of the *RF(PV) + Solar angles* model are similar or even better than those of the *RF(k3)* model. However, when comparisons between models are performed with similar inputs, the results show that the strategy based on clear-sky normalised inputs (i.e. the *RF(k3) + Solar angles* model) attains the best skill scores. Thus, this suggests that the model is not able to learn the variability patterns associated with the Sun's path from past production observations and solar angles.

Lastly, we assist the forecasting model based on non-normalised inputs by providing it with additional explicit information regarding the Sun's movement for the target time (Figure 4.23). First, the clear-sky profile is added as an additional explanatory feature, which leads to the *RF(PV) + Solar angles + CS* model. This new input has a positive impact on both accuracy scores. Yet, this approach only slightly outperforms the *RF(k3) + Solar angles* model for the highest horizons considered. Second, we replace the clear-sky profile feature with predictions of the Surface Solar Radiation Downwards (SSRD). The approach based on raw inputs (i.e. *RF(PV) + Solar angles + SSRD*) is outperformed by the method that normalises the irradiance-related inputs by the clear-sky irradiance (i.e. *RF(k3) + Solar angles + k(SSRD)*). Performance differences are significant when considering the nMAE criterion, but low with the nRMSE. With this last result, the Sun's patterns are directly carried by the SSRD predictions; however, the model based on non-normalised inputs is outperformed by its counterpart fed with clear-sky normalised data. This suggests that the normalisation process through spotlighting cloud-induced variability

of the production signal is beneficial even for advanced forecasting algorithms.

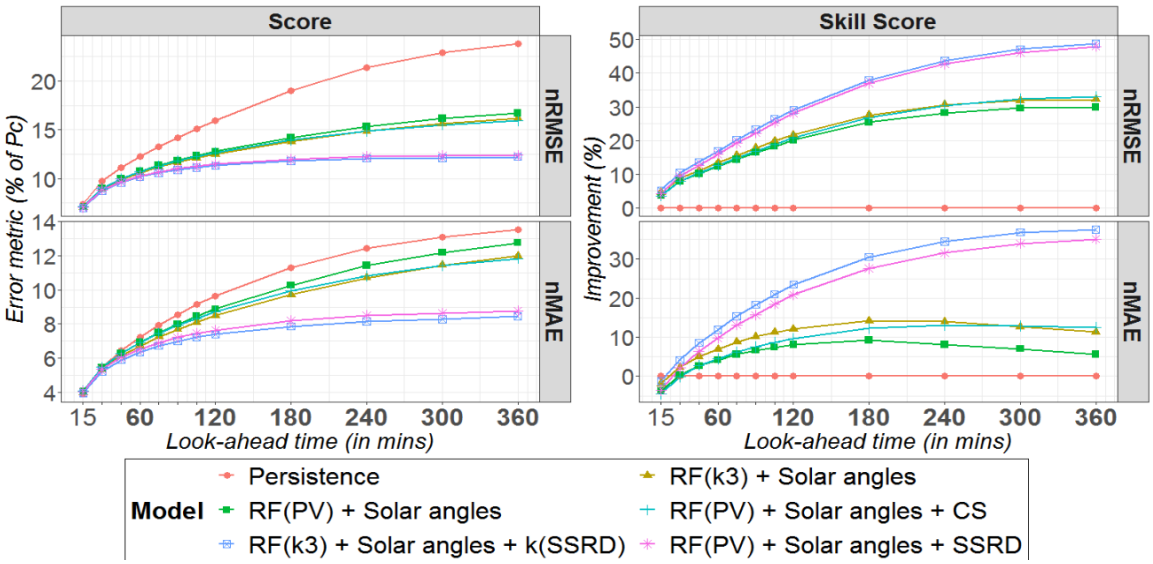



Figure 4.23 – Forecasting performances of an RF model fed with clear-sky normalised data (i.e. *RF(k3) + Solar angles*, *RF(k3) + Solar angles + CS*, *RF(k3) + Solar angles + k(SSRD)*) or non-normalised inputs (i.e. *RF(PV) + Solar angles*, *RF(PV) + Solar angles + CS*, *RF(PV) + Solar angles + SSRD*).

**Research Answer - *Normalisation process/ML (1/2)***

In this section, we investigate the influence of clear-sky normalisation over forecast accuracy. In other words, the main objective is to determine whether an ML-based model is able to learn Sun-related variability patterns on its own. When the model considers only non-normalised power observations, its accuracy is lower than when the same model is fed with clear-sky normalised inputs. Thus, it is possible to improve the performances of the model by assisting it to assess Sun-related variability patterns. To do so, we include the following explanatory features: the elevation and azimuth angles of the Sun, the clear-sky profile, and irradiance predictions. Despite the different options investigated, comparisons performed between the model fed with similar inputs reveal that normalising irradiance-related features through clear-sky irradiance model outputs leads to the best forecast accuracy.

**Research Answer - *Normalisation process/ML (2/2)***

In this section, the RF model is considered. As a complement, a similar study is carried out with an ANN in Appendix B.2. Similar conclusions are drawn, except that this model makes better use of irradiance predictions to the extent that strategies based on clear-sky normalised or non-normalised inputs provide very close accuracy scores. In a nutshell, the ANN model does not require clear-sky normalisation when irradiance forecasts are included. However, we advise forecasters to resort to clear-sky normalisation to facilitate the regression process and to achieve the highest accuracy for a wide range of configurations. In particular, the normalisation process should be used in an ST context to prevent the assimilation of Sun-induced irradiance variations with cloud movements.

Thereafter, a variable $X$ normalised by the output of a clear-sky model is simply denoted as $\bar{X}$.

#### 4.4.2.3 Temporal dependency

A time series may be auto-correlated, that is to say, the instance at time $t$ may be correlated with previous or following observations. Then, the question arises of which number of lagged terms to include in the forecasting model. To investigate the time dependency of CSI, the ACF and Partial Auto-Correlation Function (PACF) tools are used. Such tools are widely used to identify the orders p and q (number of time lags) of an ARMA model.

On the one hand, the ACF at lag $k$ computes the correlation between variable $Y_t$ and its delayed copy $Y_{t-k}$. This enables us to find repeating patterns in the time series. In other words, the ACF measures the similarity between observations as a function of the time lag, as such it can be seen as a rough measure of the ability to forecast the time series at time $t$ from previous observations. On the other hand, the Partial Auto-Correlations (PAC) at lag $k$ is the correlation measured after removing the effect of any correlations due to the terms at shorter lags. A PAC which significantly differs from 0 indicates lagged terms that are useful predictors of the feature $Y_t$.

The correlogram (i.e. the left part of Figure 4.24) shows that the ACF exponentially decreases as the lag $k$ increases, while the PAC associated with the first 9 lags are significantly different from 0. According to Table 6.1 from [234], this characterises an AR process of order 9. Given the high PAC observed for the very first lag, it could have been adequate to resort to an AR(1) process, yet, we chose to add previous lags for the diversified information they carry and let the model's feature selection algorithms select the most relevant inputs. In a nutshell, Figure 4.24 reveals that significant information is still present until the $9^{\text{th}}$ lagged term. The vector $\{X_{t-135min}, \ldots, X_{t-15min}\}$ of previously observed CSI is then used as input to predict the value of $y_{t+h}$.

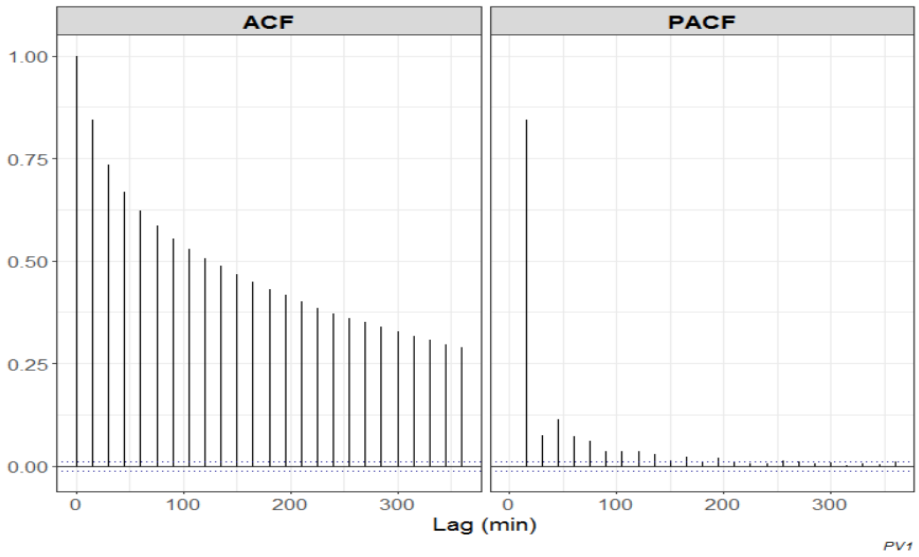



Figure 4.24 – ACF and PACF of the complete time series of normalised power (i.e. $k_t^3$) observed during year 2016 but excluding periods with zero clear-sky irradiance. The blue dotted line shows the zone outside which $k_3$ has statistically significant correlations with its historical values.

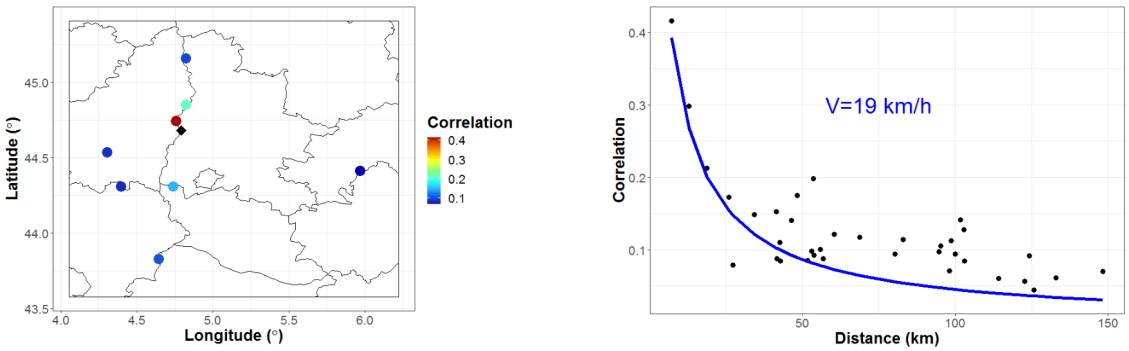
#### 4.4.2.4 Spatial dependency

The main idea of the present section is to highlight the spatial dependencies that may exist among PV production sites and with satellite-based inputs. We can estimate the correlation between pairs of stations with the Cross-Correlation Function (CCF). The latter is a measure of the similarity between two time series as a function of the displacement (lag) of one relative to the other. In this paragraph, we consider that the displacement is null in order to focus only on the spatial correlation between features. To measure the spatial dependence between features we adopt an approach inspired from [84, 235]. The focus is on the changes in the CSI ($\Delta k_t = k_t - k_{t-1}$) rather than on the CSI. Such an approach avoids falsifying the correlation by removing the remaining seasonal effects. This is in line with conclusions from the previous section, which states that CSI needs differentiation to achieve stationarity. Then, the cross-correlation between the changes in the CSI of two time series is computed.

In a first step, we investigate the spatial correlations that may exist between stations in the fleet. Figure 4.25a reveals that PV6 possesses the highest correlations with its northern and southern neighbours. This highlights that close sites are more related than distant ones. However, given the north/south configuration of the network, it is difficult at this stage to quantify the direction and strength of the spatial relationships. It is worth mentioning that the correlation levels observed here differ from previous works (e.g. [236]) owing to the fact that ($\Delta k_t$) is considered instead of ($k_t$). Figure 4.25b shows the exponential decay of the station pair correlation as a function of station distance. In [237], the authors provide a review of correlation formulations as a function of the station distance ($d$), the sampling time ($\Delta t$) and the prevailing regional cloud speeds ($V$). The experimental correlations in

Figure 4.25b correspond well with Equation 4.19 which considers a regional prevailing cloud speed of 19 km/h.

$$\rho = \frac{1}{1 + \frac{d}{\Delta t \cdot V}} \tag{4.19}$$




(a) Spatial correlation between PV6, represented by the black dot, and the other power plants.

(b) Spatial correlation of the station pairs as a function of site-pair distance. Blue line represents Equation 4.19 with $V = 19$ km/h.

Figure 4.25 – Yearly spatial correlations between power unit observations.

Then, we move our focus on spatial correlations between production time series and satellite-based irradiance observations. As for the distributed network of power units, we observe a rapid decay in the correlation around the position of the plant of interest (Figure 4.26). This stresses the relevance of SDSI observations to account for the PV production variations. The low value of the correlation may be explained by the difference in the spatial resolution of both sources of information.

#### 4.4.2.5 Spatio-temporal dependency

In a next step, to complete the correlation analysis it is necessary to assess the temporal correlations between pairs of power units. To do so, we analyse the CCF for different values of the displacement (or "lag") term.

At a restricted temporal scale (i.e. one day), Figure 4.27 clearly depicts that two distant sites may be affected by the same cloud effects with a temporal lag which depends on the distance and the atmospheric structures' velocity. Here, a phase difference of around 15-minutes is quantified by computing the daily temporal lag between two stations that maximises the cross-correlation of the two $\Delta k_t$ time series. By considering the daily cross-correlation we implicitly assume that cloud propagation remains consistent the whole day.

This method is then applied iteratively on yearly data. Figure 4.28a represents the distribution of the daily temporal lag between two stations. The highest occurrences of time lags gather between $[-30min; +30min]$, which assumes that these two power units are affected by similar events with about 30 minutes delay (which is plausible given the low distance between the units). We observe a high frequency of inter-correlations reaching

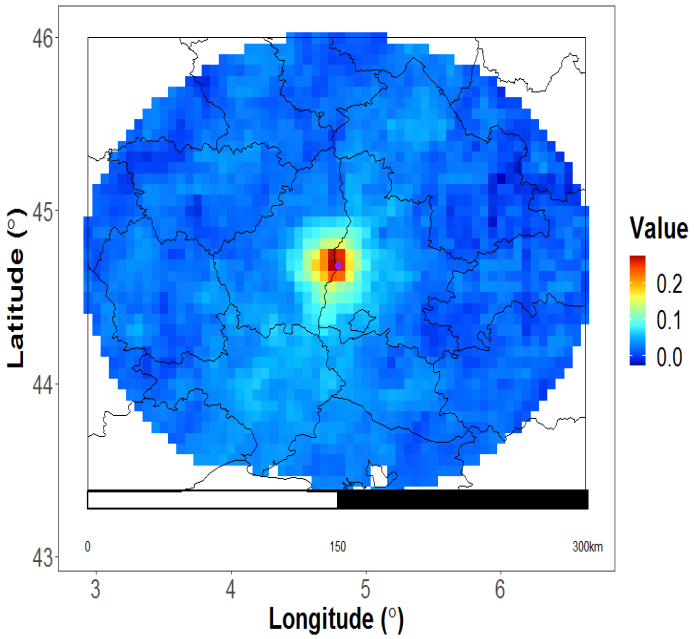



Figure 4.26 – Cross-correlation between production observations of PV6 and satellite-based information. The central purple point stands for the position of the power unit.

their maximum values for a null temporal lag. This may be explained by the fact that close sites experience similar weather conditions (e.g. overcast or sunny days). In a next step, the distribution of the time lag maximising the daily cross-correlation is computed for all the station pairs. The Hovmoller diagram displayed at Figure 4.28b shows the temporal shift distribution for all possible pairs of power units as a function of the distance. As the separation distance between station pairs increases, the magnitude of possible time lags between two power units (i.e. the range of temporal lags associated with the highest frequencies) increases. This illustrates metaphorically that "the further we look into time, the further we look into space": as the lead time of forecasts extends, so does the distance of plants used as explanatory features.

Lastly, ST dependencies between production observations and the SDSI are assessed. In Figure 4.29, the propagation time of weather structures is computed between the PV power unit and all the pixels of the satellite-based image as the temporal lag that maximises the cross-correlation coefficient. Interestingly, this figure shows propagation from the eastern longitudes towards the western longitudes. At first glance, this analysis refutes the wind distribution observed in Figure 4.30a. The latter shows a dominant stream coming from the northern direction, and a secondary prevailing stream coming from the south. The dominant stream is associated with the Mistral, which is a strong, cold wind accompanied by clear cool weather. As such, the Mistral does not play a significant part in the transport of clouds. On the contrary, the south wind is associated with moisture-laden air coming in from the sea that causes cloud cover. These prevailing north and south distributions can be accounted for by the effects of topographic relief of mountains. Indeed, the Rhone valley is characterised

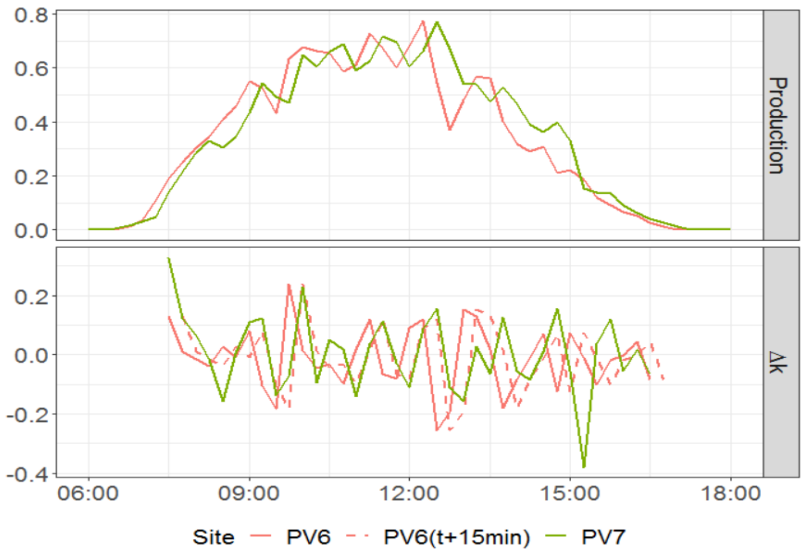
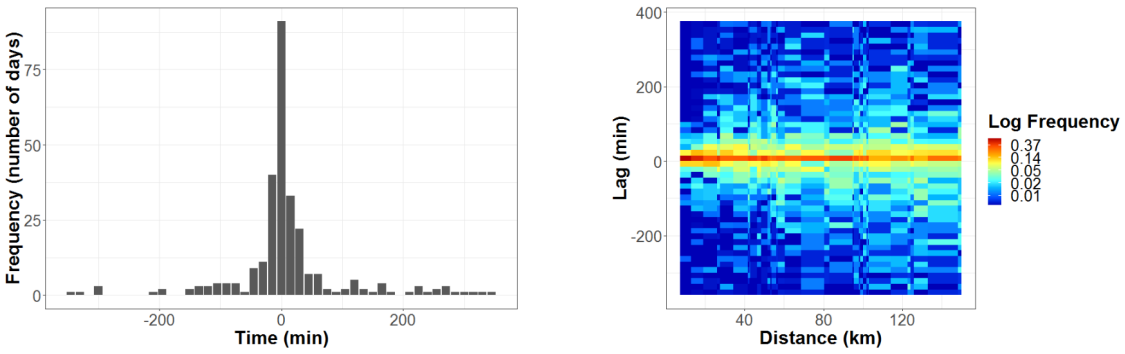



Figure 4.27 – Production observations and associated changes in the CSI of PV6 and PV7 (18.91 km apart) on 2016-02-26. The production series are displayed in the top window while the change in the CSI ($\Delta k_t$) is shown at the bottom. A maximum correlation is attained for time $t + 15min$. The dotted line represents the 15-min lagged values of PV6.




(a) Distribution of the time lag maximising the daily cross-correlation between change in CSI of PV6 and PV7 (18.91 km apart).

(b) Hovmoller diagram [238] representing the propagation of the daily distributions of time lags between pairs of stations as a function of distance.

Figure 4.28 – Distribution of time lags obtained with the cross correlation function. Only values higher than the 95% confidence interval are retained.

by a funnel-shaped relief oriented along the north/south axis (Figure 5.9). At the 850 hPa pressure level, we are still in the low layers of the atmosphere, which explains the orography dependence of winds. At higher layers such as 500 hPa (Figure 4.30b), the influence of the oceanic air flux (i.e. westerly winds) becomes significant. Clouds associated with this flux are often more scattered than clouds associated with northward winds. Therefore, ST structures advected by western winds are often more numerous. This explains the power production dependencies of westerly weather structures.

Based on extreme temporal lags observed at the western edge of the correlation area, which are roughly between −150-min and −100-min, a simple calculation gives an atmo-

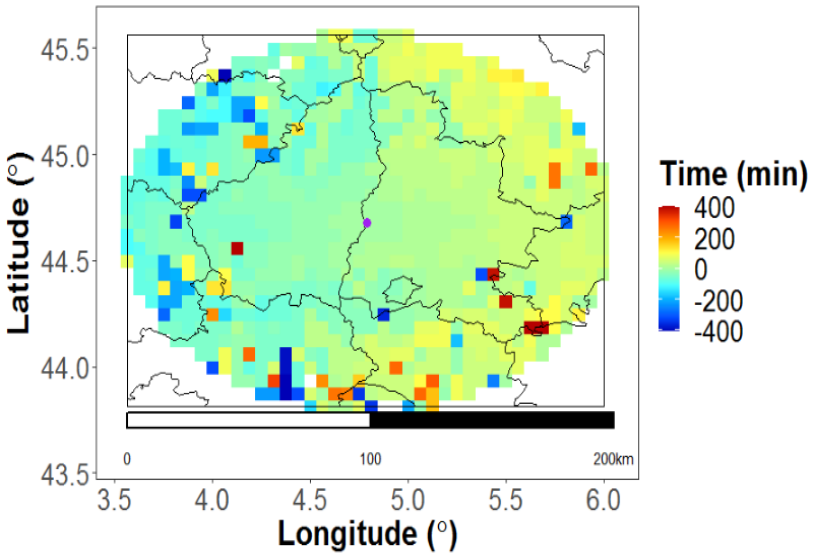
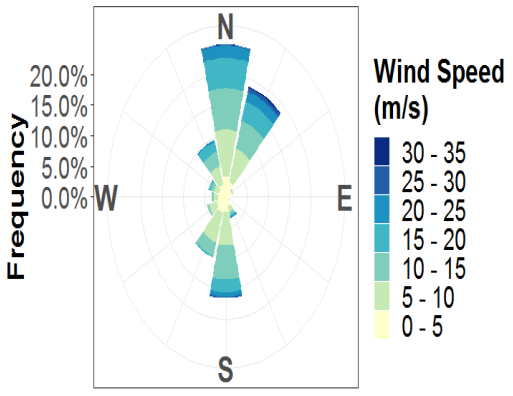
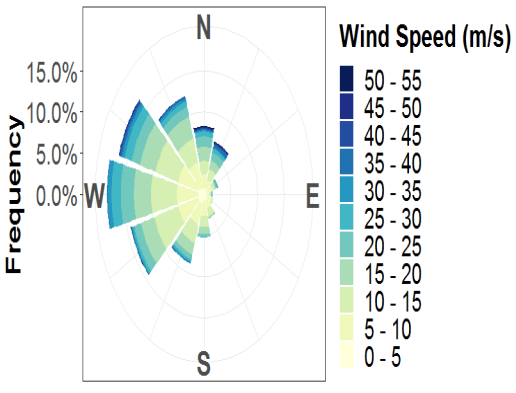



Figure 4.29 – Propagation time from the PV farm marked by a purple point. Propagation time is associated with the maximum cross-correlation between normalised values of production and SDSI.

sphere structure speed displacement of around $40 - 60$ km/h (i.e. $11.11 - 16.66$ m/s). This value is consistent with the western wind speeds observed in the wind rose (Figure 4.30b). Such values are significantly distinct from the values derived at Figure 4.25b. These gaps may be explained by the fact that Figure 4.29 exhibits a westerly tendency of weather structure displacement, while power plants are mainly located in the north-south axis, which may invalidate Equation 4.19. It is worth mentioning that the observed predominance of east winds does not result from the Sun's path, and so a poor clear-sky normalisation approach.




(a) Wind speed and direction distribution at 850 hPa. Cumulus clouds usually form at this pressure level [62]. Mean wind speed is 9.81 m/s.




(b) Wind speed and direction distribution at 500 hPa. Mean wind speed is 14.11 m/s.

Figure 4.30 – PV6 wind speed and direction histograms at 850 and 500 hPa for year 2015.

This analysis highlights the propagation of ST dependencies and confirms the interest

of forecasting solutions based on temporal and spatial relationships among inputs.

## 4.5 Conclusions

The main objective of this chapter was to investigate the integrity of the datasets at our disposal, while analysing and exhibiting the dependencies between the different features.

In a first step, an identification and imputation strategy is developed to deal with spurious power observations. Actions taken within this chapter highlight that in the presence of a dataset contaminated with fallacious observations it is better to reject them, even though the literature shows us that missing data tends to degrade forecasting performances (the results were obtained within the wind power forecasting field, but it is assumed that such conclusions can be extended to the PVPF field). The corrective approach developed in this chapter shows that forecasts performed on datasets in which fallacious production have been either removed or corrected exhibit rather similar accuracy. This behaviour may be explained by the fact that: (1) the degradation induced by missing data is compensated by errors generated by the corrective process due to the limited quality of proxy observations, or (2) production malfunctions are associated with a MCAR process, which does not alter the data distribution, and consequently has no impact on model estimations. To extend the conclusions drawn in this chapter, it could be interesting to compare the developed corrective approach with classic imputation strategies found in the literature such as tree-based or mean-based imputation methods. These statistical methods differ fundamentally from the one investigated here, which is derived from physics-based principles.

The clear-sky normalisation process is implemented to explicit the information contained in the inputs. A performance- and test statistics-based comparison highlights that normalising power observations with the clear-sky power leads to slightly higher performances. In the ML-related literature, it is common to assume that models are able to account for signal seasonality on their own, yet we highlight that such regression tools may benefit from the normalisation process. Lastly, we demonstrate the presence of ST relationships between inputs using the correlation coefficient.

## 4.6 Résumé en Français

Comme nous avons pu le constater au travers du précédent chapitre, la chaîne de conversion de l'irradiance en électricité est longue et complexe, si bien que le système d'acquisition de données est à même de subir diverses avaries et de remonter des données incomplètes ou corrompues. Ce phénomène détériore la qualité du signal de production et a des répercussions sur la précision des prévisions. Il convient alors d'identifier les données fallacieuses et d'appliquer un traitement adéquat. En plus de contenir une composante liée aux avaries techniques, le signal de production est également constitué de deux autres composantes : l'une résultante du mouvement du soleil (qui induit une variabilité journalière et saisonnière)

et l'autre associée aux mouvements de masses atmosphériques. Dans un contexte d'utilisation de méthodes de prévision ST, la composante déterministe liée à la course du soleil peut altérer les dépendances entre plusieurs points d'observations spatialement distribués. En effet, lorsque les stations sont alignées selon un axe Est-Ouest, ces dernières ont tendance à montrer une plus forte corrélation, non pas en raison d'une plus grande dépendance aux mouvements des masses atmosphériques mais plutôt en raison du déplacement du soleil dans la voute céleste. Ainsi, ce chapitre a pour objectif de proposer une méthode permettant d'éliminer les sources de variabilité qui viennent polluer le signal de production. En d'autres termes, l'objectif est d'affiner les données afin de mettre en lumière l'information pertinente.

### Identification et correction des données aberrantes

Dans un premier temps, notre questionnement portant sur la fiabilité des données de production nous a conduit à remettre en question les informations disponibles concernant la géométrie des centrales (i.e. angles d'inclinaison et d'orientation des modules). En effet, bien souvent l'installation des modules est dictée par la topographie du terrain. Lors des premiers temps de ce travail de recherche, une approche naïve avait été proposée pour déterminer des angles plus en accord avec la réalité du terrain. Néanmoins, la faiblesse de la méthode et les importantes déviations vis-à-vis des valeurs standard nous avaient conduit à préférer les valeurs par défaut. Désormais avec le recul, il pourrait être pertinent d'identifier des journées de production ciel clair, puis d'estimer la production théorique en condition ciel-clair à partir des sorties de modèles ciel-clair et en utilisant le modèle de conversion proposé au Chapitre 3. L'idée serait ensuite de retenir les angles conduisant aux plus faibles écarts entre les deux jeux de données.

La suite de notre investigation consiste à identifier et éventuellement corriger les données qui s'écarteraient d'une caractérisation fidèle de la réalité. Les défauts techniques peuvent altérer les données ou même induire une perte d'information, ce qui en aucun cas ne reflète la réalité des choses. Dans ces conditions, on comprend qu'il devient difficile d'établir des relations statistiques pertinentes lorsque des données exogènes telles que des variables NWPs, qui sont exemptes de tels défauts, sont utilisées. Un contrôle préliminaire de la qualité des données est tout d'abord réalisé en nous basant sur un contrôle générique des séries temporelles :

1. Vérification globale de la série temporelle : absence de changements d'heure été/hiver, absence d'observations associées à une éclipse solaire, absence de données supérieures à la capacité installée.
2. Contrôle rapide de la qualité des observations : e.g. les jours pour lesquels le ratio entre production observée et production théorique ciel clair est très faible sont identifiés comme problématiques.
3. Contrôle des données constantes : on observe une production constante sur plusieurs instances temporelles lorsque la production est bridée ou la centrale à l'arrêt.

Les critères utilisés lors de cette étape préliminaire de validation sont aisés à implémenter mais ne permettent pas d'identifier des avaries associées à des arrêts de sous-composants de la chaîne de conversion (e.g. onduleur ou transformateur). L'objectif et la contribution phare de ce chapitre résident dans le développement d'une stratégie d'identification et de correction des données de production aberrantes. Afin d'identifier avec certitude un comportement déviant, nous utilisons un proxy de la situation météorologique sur site, en l'occurrence (1) des observations de l'irradiance sur site et (2) des mesures au niveau des transformateurs. Une première approche basée sur des prévisions NWPs et des observations satellite de l'irradiance n'a pas été concluante en raison de la résolution plus ou moins grossière des données et des erreurs d'observations et/ou de modélisation. Dans la mesure où les données que nous avons retenues ne sont actuellement pas valorisées parCNR, aucun suivi de l'instrumentation n'est réalisé, si bien que ces dernières sont entachées d'erreurs. Les données ont donc été analysées afin d'identifier, par exemple, des phénomènes de données gelées ou des décalages temporels.

Pour les deux types de données, des approches spécifiques ont été développées. Dans un souci de concision, la suite de ce résumé est dévolue à la méthode basée sur des observations d'irradiance à partir de cellules de référence. Une comparaison visuelle sur plan 2D, où chaque axe représente respectivement la production observée et la production simulée à partir des observations et du modèle de conversion développé au Chapitre 3, met en lumière plusieurs régimes de production. La méthodologie développée consiste tout d'abord à segmenter les séries temporelles (typiquement en un ou deux jours). Ensuite, ces segments sont transférés dans un nouvel espace construit à partir de la variable cible (i.e. les observations de production) et la variable utilisée comme proxy (i.e. la production simulée). A l'intérieur de cet espace, l'algorithme de clustering DBSCAN permet de regrouper les différentes instances en des groupes partageant des spécificités communes. Ces groupes sont ensuite associés à un régime de production et un facteur de correction est appliqué sur les observations fallacieuses. Nous montrons par la suite que ce processus est fortement tributaire du site étudié, ceci s'explique par des architectures plus ou moins complexes, et par la qualité des données d'entrée utilisées comme proxy. Quoi qu'il en soit, cette approche permet d'améliorer le coefficient de détermination entre les observations et les simulations de la production.

Pour évaluer l'impact du processus développé ci-dessus, nous considérons successivement (1) un jeu de données brutes, (2) un jeu de données pour lequel les données fallacieuses sont rejetées (i.e. remplacées par la valeur NA), et enfin (3) un jeu de données où les observations identifiées comme étant erronées sont corrigées. Ces différents ensembles de données sont ensuite utilisés pour entraîner un modèle AR et un modèle RF. L'analyse des performances se fait quant à elle sur des données corrigées afin de se rapprocher le plus possible de la réalité météorologique. Nous observons pour le modèle AR que, quels que soient les horizons considérés, les deux stratégies de rejet et de correction des données conduisent à une

amélioration de la nRMSE par rapport à un modèle entraîné sur des données brutes. Les conclusions concernant la nMAE sont moins tranchées : nous observons une dégradation du score pour les horizons inférieurs à 120 minutes. Cela signifie que le modèle calé sur les données brutes à tendance à faire moins d'erreurs faibles. Un comportement similaire est observé pour le modèle RF. Dans ce cas, néanmoins, les variations de performances par rapport au modèle de référence sont moins prononcées. A notre étonnement, nous constatons que le recours aux données corrigées n'améliore pas significativement la précision des prévisions comparativement aux modèles générés à partir des données exemptes d'observations fallacieuses. Ceci pourrait s'expliquer par le fait que la perte d'information générée par les données manquantes est contrebalancée par le bruit introduit lors de la phase de correction.

### Mise en évidence de l'information pertinente

La dernière partie de ce chapitre a pour but de mettre en valeur l'information pertinente contenue dans le signal de production, ou les variables liées à l'irradiance d'une manière générale. Pour ce faire, nous nous tournons vers le processus de normalisation qui consiste en la division du signal étudié (e.g. la production PV, ou l'irradiance solaire) par une grandeur analogue théoriquement observée sous un ciel dépourvu de nuage. La valeur qui en découle se nomme alors l'indice ciel-clair. Cette approche permet de mettre en lumière la composante stochastique liée à la variabilité des masses atmosphériques et de faire fi de la composante associée à la course du soleil. Une étude bibliographique a permis de mettre en lumière les spécificités des différents modèles présents dans la littérature. Notre choix s'est porté sur le modèle McClear.

Les variables en lien avec l'irradiance sont par nature non-stationnaires. En un mot, la stationnarité signifie que les propriétés statistiques d'une série temporelle ne changent pas au cours du temps. Cette propriété est souvent un prérequis pour l'utilisation de nombreux modèles de régression de série temporelle. Dans la littérature dévolue à la prévision de la production PV, la normalisation par l'irradiance ciel-clair sur plan horizontal est souvent utilisée comme moyen de stationnarisation. Nous mettons en évidence que, formellement, cette approche ne permet pas d'obtenir des séries stationnaires. Néanmoins, la considération de l'irradiance ciel-clair sur plan incliné ou même la puissance électrique associée permet d'atteindre de meilleures propriétés de stationnarité locale et d'améliorer les performances prédictives.

Dans la littérature consacrée à l'utilisation de modèles d'apprentissage machine (ML), le recours à des grandeurs brutes (dans le sens où celles-ci ne sont pas non normalisées par la sortie d'un modèle ciel clair) est souvent de mise. Ceci peut s'expliquer par la croyance en les capacités du modèle à intuiter par lui-même les mouvements du soleil. Dans les faits, nous montrons qu'un modèle RF alimenté par des variables normalisées par des sorties de modèle ciel clair atteint des performances supérieures vis-à-vis du même modèle calé sur des variables non-normalisées. Il est possible d'aider le modèle en lui fournissant des informations

supplémentaires telles que les angles solaires (i.e. angles d'élévation et d'azimut) ou le profil ciel clair.

### Dépendance des différentes variables étudiées

Enfin, dans la mesure où la normalisation ciel clair permet de supprimer, ou tout du moins, de réduire l'influence du mouvement du soleil, il a été possible de mettre en évidence les dépendances ST pouvant exister entre les différents sites de notre cas d'étude ou avec des données issues de l'imagerie satellite. Une analyse de la distribution des différents régimes de vent au niveau des sites d'intérêts permet d'expliquer les dépendances observées.

# Chapter 5

# Spatio-temporal Information

> Many a trip continues long after movement in time and space have ceased.
>
> John Steinbeck, *Travels with Charley: In Search of America* (1962)

**Contents**

## 5.1 Introduction

With the growth of Photovoltaic (PV) energy, new production facilities are flourishing just about everywhere. This development, combined with advances in smart monitoring and measurements, paves the way for a paradigm shift in PV power forecasting from temporal- to Spatio-temporal (ST)-based forecasting models. ST methods assume that weather features exhibit correlations among close-by areas with temporal lags depending on the spatial distance between sites and the propagation speed of weather structures. This new paradigm offers power producers the possibility to economically value information from geographically distributed solar plant networks in the form of forecast accuracy improvements due to correlations between units, and prepares the ground for a data-sharing market [60].

Distributed PV production observations have been used in the literature on solar energy short-term forecasting for a couple of years due to their proven advantages over temporal forecasting methods [239], but they are still investigated in the light of new modelling strategies. In 2015, [83] presented an ST forecasting method based on the Vector Auto-Regressive (VAR) framework combining observations of solar generation collected by smart meters and distribution transformer controllers. Recently, [240] proposed an ST deep learning framework. In line with what was performed in Section 4.2, this information can also be used to address partially corrupted observations in training datasets. In this regard, [241] proposes a co-kriging strategy to complete data points for which data are not available based on ST dependencies between sensor observations. Some preliminary studies performed as part of this research have shown that for very close sites, ST dependencies could be used for missing entry imputation. This avenue has not been further investigated because it cannot be generalised to our whole PV unit network.

Depending on its distribution or density, a PV network may partially account for the complex ST processes at stake (e.g. mainly sites located upwind or crosswind). To fill these gaps, satellite-based observations are an appealing option. With recent developments, geostationary satellites can capture images of Earth at a temporal resolution of less than an hour, which enables operational uses. Aguiar et al. in [55] demonstrate the positive impact of Satellite Derived Surface Irradiance (SDSI) observations on forecasts based only on endogenous features for intra-day solar forecasting. In this study, Artificial Neural Networks (ANN) are fed with past endogenous observations and a subset of 30 pixels obtained from the Pearson correlation-based selection. In addition to traditional satellite-derived features found in the literature, we also consider opacity maps obtained from infrared channels. Despite being under-represented in the literature (only two studies have been found [104, 114]), infrared channel-based data present the advantage of offering nighttime observations, which contributes to improve early morning forecasts.

**Research Gap - *Mixing of data sources***

Traditionally in the literature, distributed PV production and satellite-derived observations are used separately. Yet both carry different information due to their spatial resolutions (i.e., distributed units enable the observation of smaller cloud structures). As a result, we might wonder whether used together, these inputs could contribute to forecast accuracy.

Therefore, in this chapter, the focus is on the use of several sources of information rather than on developing a dedicated forecasting model dealing with ST datasets. In this regard, techniques are investigated to make the most of available information, while reducing the computational burden inherent to spatial observations. The general workflow of this chapter is displayed in Figure 5.1.

## 5.2 Spatially distributed PV plants

### 5.2.1 State-of-the-art

A variety of techniques can be applied to generate forecasts from ST observations. For instance, kriging [238] refers to a group of geostatistical weighted interpolation methods based on ST statistical dependencies, with weights derived from co-variances or correlations among observations of a random field. These methods are widely used in the domain of spatial analysis, and are currently applied to the solar and PV generation forecasting fields (e.g. [85, 242, 243]). They enable predictions at unobserved locations, which offers the possibility of quantifying the PV potential of new projects. One major drawback of kriging is that it is computationally intensive. In this regard, [243] proposes a method to quantify the spatial and temporal decorrelation distances (i.e. spatial and temporal distances from which two sites are uncorrelated) to reduce the size of the problem (number of sites and/or time-steps). In addition, it requires special care regarding the ST structure modelling and the associated choice of covariance/correlation function. In [85], the authors highlight that applications of kriging methods go far beyond the field of solar energy, which implies that ST covariance functions may not model irradiance effectively. Lastly, kriging methods are efficient with rich ST data, but not for a reduced number of measurement sites [84], which can be an issue regarding the density of the PV unit network under study.

Computing correlation metrics between several PV power plants or satellite pixels can be intense. Here, the option is to consider data-driven frameworks (i.e. the Auto-Regressive (AR) and Random Forest (RF) models) that are not initially designed for forecasting tasks involving ST correlations, but make it possible to easily consider ST features as additional exogenous variables. A large part of the ST-related literature is dedicated to Auto Regressive Integrated Moving Average (ARIMA)-based models due to their capability of including

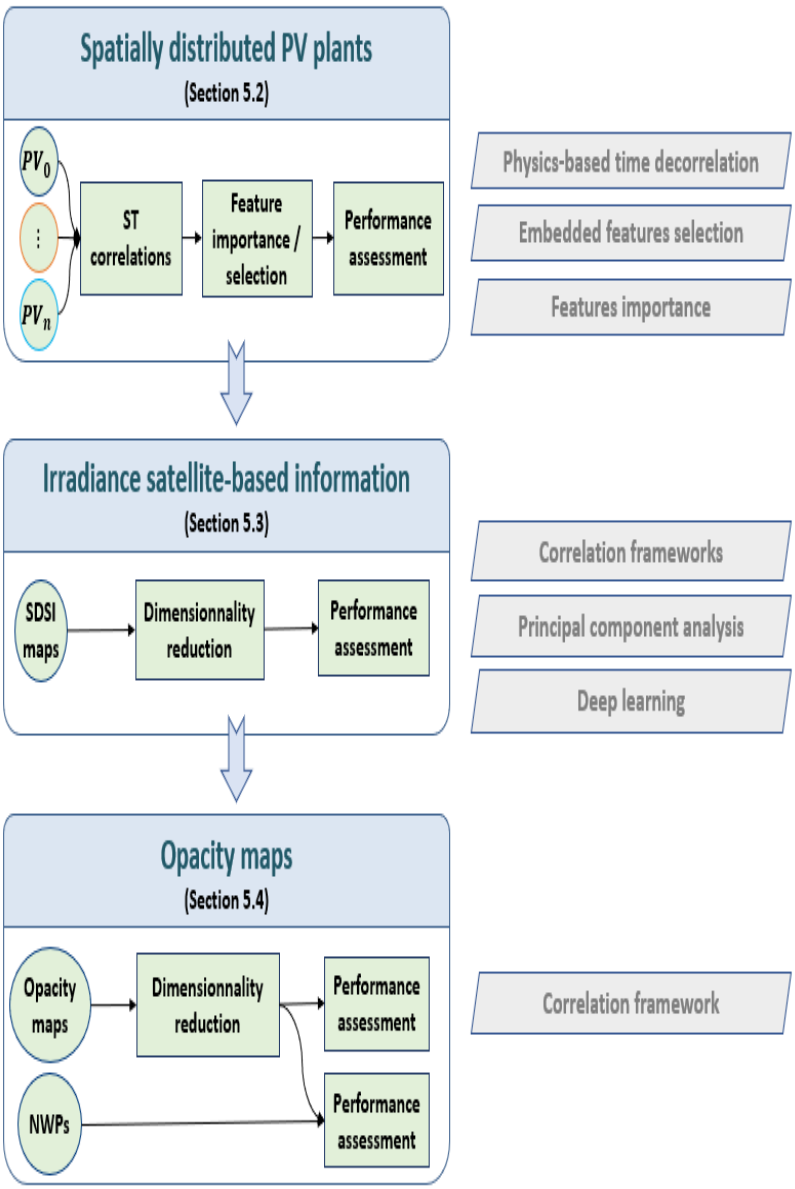



Figure 5.1 – General workflow of the chapter.

numerous data from different sources [56, 58, 62, 83, 244]. In 2015, Bessa et al. [83] were among the very first researchers to present an ST forecasting method based on the VAR architecture, combining observations of solar generation collected by smart meters and distribution transformer controllers. The results show that ST methods outperform temporal methods for all the timescales under study. The results indicate that information from distributed PV generation can reduce the forecast error by between 8% and 12% on average for the first three lead times compared to an AR model, and that improvement deteriorates with the lead time. Later, in 2018, Agoua et al. [58] proposed a new normalisation technique to overcome weaknesses of the Clear-Sky Index (CSI) for early and late hours of the day. To take into account the local weather conditions in the ST model, the authors propose an AR framework with weather-dependent coefficients. Compared to the ST model with

fixed parameters, this model conditioned on wind speed shows a deteriorating reduction in the Root Mean Square Error (RMSE) for the first two hours: for the 30-min horizon, the improvement reaches around 2% and becomes non-existent beyond 2.5 hours. In the same vein, in 2019, Amaro e Silva et al. [62] developed a wind regime-based approach, where different Auto-Regressive with eXternal inputs (ARX) models are trained for different wind speed and direction intervals within the scope of very short-term solar forecasting. This regime-based approach detects different spatial patterns achieving skill scores greater than 20%. Tree-based approaches are also represented in the ST context. Huang et al. [245] propose a comparison between several data-driven models (i.e. Gradient Boosted Regression Trees (GBRT), Support Vector Machines (SVM), ANN, ARX) with consideration of both spatial information from large-scale neighbouring sites (i.e. 65 sites within 30 km) and temporal information. The study highlights that the GBRT model gives the best performances with the lowest normalised Root Mean Square Error (nRMSE) and the highest $R^2$. The authors explain this success by the ability of the model to take advantage of both the regression trees and the boosting technique so that it is insensitive to outliers, flexible enough to express solar data features, powerful enough to fit complex nonlinear relationships, and capable of performing automatic feature selection. Nevertheless, no comparison between temporal and ST modelling is provided. Persson et al. [246] investigate the potential use of GBRT for short-term solar power generation forecasting in a multi-site framework composed of 42 rooftop installations. The main findings are: (1.1) for the same inputs, the GBRT model outperforms a recursive AR model, (1.2) AR-based forecasts are more accurate for low-variability sites, and (2) the ST version of the model exhibits improved scores compared to their temporal counterparts. These results motivate the use of tree-based models.

### 5.2.2 Model definition

Since ST information is considered as exogenous inputs, their integration in the root forecasting models defined in Chapter 2 is straightforward. The new inputs vector is defined as follows:

$$X_t^{x,\mathsf{T}} = \begin{bmatrix} P_{t-l}^x & \dots & P_t^x & P_{t-l}^{\mathcal{S}} & \dots & P_t^{\mathcal{S}} \end{bmatrix} \tag{5.1}$$

- $x$ Site of interest, for which the forecast is performed,
- $l$ Order of the AR model (i.e. number of temporal lags considered),
- $\mathcal{S}$ Set representing the $s$ neighbours of $x$ included in the ST model.

### 5.2.3 Spatio-temporal correlations between plants

To avoid considering irrelevant sites that are too far away to have an influence over the production of the site of interest, we assess the spatio-temporal correlations that may exist between sites. The degree of correlation between features is measured though the Mutual Information (MI) criterion, which is defined later on in Section 5.3.2.1.1.

Figure 5.2 shows that as the forecast horizon extends, spatio-temporal correlations between sites decrease. A decrease in correlation is also observed with increased distance as expected. Naturally, the highest MI score is reached at the horizon $h = 00H15$ for the last production level measured at PV1 (i.e. at time $t$). Besides, PV1 and PV4 are hardly correlated independently from the forecast horizon or lag considered, while we observe that despite the distance, PV1 and PV8 exhibit interesting correlations. This is due to the fact that in the Rhone Valley dominant streams come from a northern direction, while secondary prevailing streams come from the south (Figure 4.30a). The former regime corresponds to the Mistral (i.e. strong winds accompanied by clear, cool weather that do not play a significant role in the transport of clouds), while the latter is associated with moisture-laden air coming in from the sea that causes cloud cover. Thus, plants in the north (such as PV4) do not have a prevailing impact on ST correlation in comparison with PV8 which is located in the south (the spatial distribution of plants is displayed in Figure 2.5). Surprisingly, we observe that for the same horizon, the correlation level between PV1 and and the last observation of PV10 is not negligible. As both sites are separated by nearly 100.00 km, it is hardly conceivable that they experience similar atmospheric conditions 15 minutes apart. Based on this observation, and the analysis performed in Section 4.4.2.5 from Chapter 4, we consider data observed within a threshold distance of 80 km.

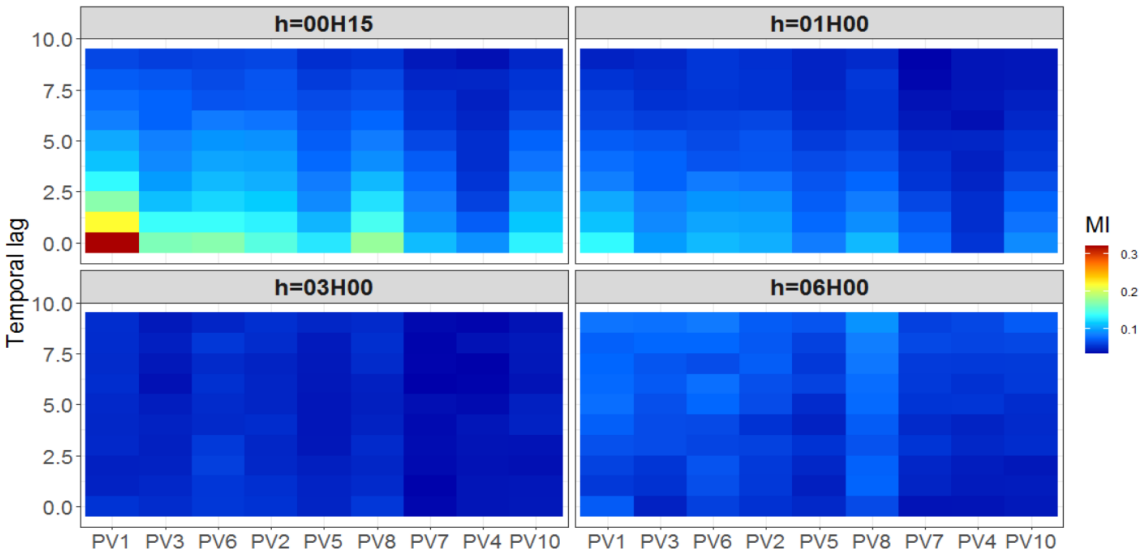



Figure 5.2 – MI between the differentiated CSI (i.e. $\Delta k_t$) of PV1 at time $t + h$ with the lagged $\Delta k_t$ of the different power plants. Sites are ordered as a function of pairwise distance. We consider the 9 previous lags.

### 5.2.4 Limits of embedded feature selection algorithm

A straightforward injection of Spatially Distributed Units (SDU) features in the regression models may lead to counter-intuitive behaviours.

#### 5.2.4.1 Performance degradation

Figure 5.3 represents the forecasting performances of different configurations of models based on ST features (observations of neighbouring sites within an 80 km threshold distance and their temporal lags) w.r.t. the same model based only on temporal information at the site of interest. This graph demonstrates that (1) in a general way, the consideration of neighbouring plants improves the forecast accuracy (the *RF + SDU(t-9:t)* model is generally better than the *RF* model), except (2) on the very first horizon where the *RF* model leads to better performances in terms of normalised Mean Absolute Error (nMAE).

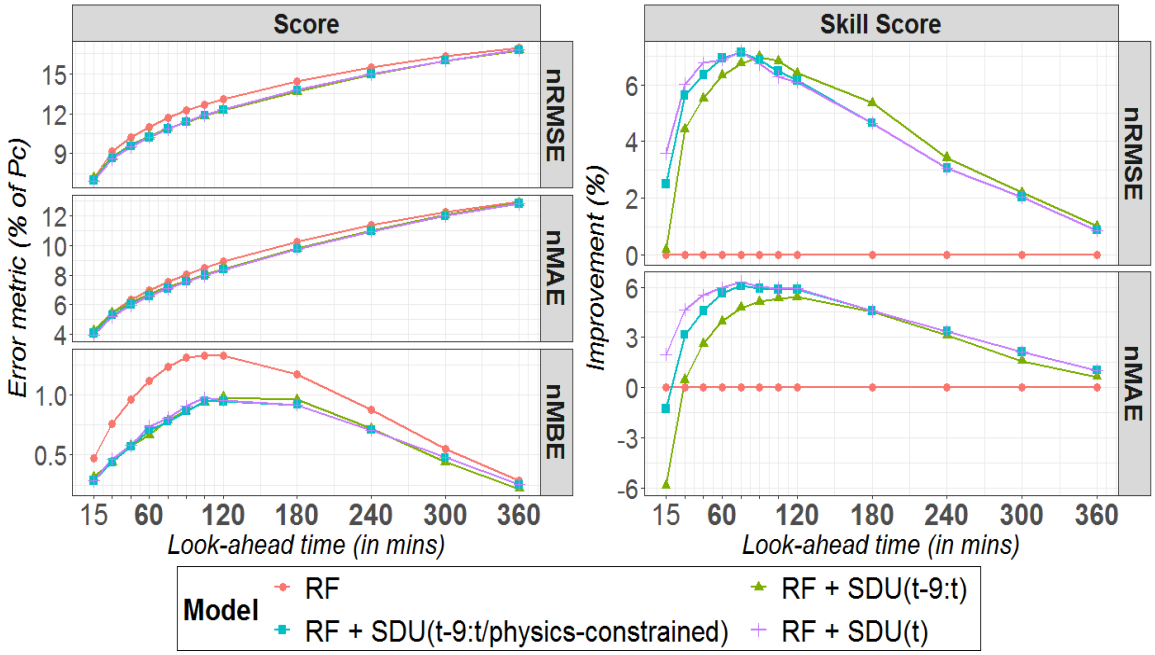



Figure 5.3 – Forecasting performances of RF models considering temporal and ST information. The lags of neighbour observations are indicated in parenthesis. The physics-constrained feature selection refers to the methods introduced in Section 5.2.4.2.

Thus, the embedded feature selection algorithm of the RF model is not able to efficiently select relevant features in an ST context for the very short-term horizons. Similar behaviour has already been observed in the literature. The author in [244] notes that an extra preselection step is useful to remove bad predictors. The latter, combined with the Least Absolute Shrinkage and Selection Operator (LASSO), improves forecast accuracy compared to a single-stage LASSO, even in the case of a low dimensional dataset (namely 17 radiometers). This motivates us to consider features preselection not only in space, but also in time.

#### 5.2.4.2 Physics-based time decorrelation distance

To reduce the number of inputs considered during the regression process, a physics-constrained feature selection approach is studied. The idea is to consider the propagation time of ST information from one site to another, derived from wind speeds, to select relevant

input lags. It is worth mentioning that [244] proposes an option based on the Mueen's algorithm for a similarity search, which does not require any prior meteorological information.

First, the bearing [1] between the site of interest, $x$, and its neighbour, $i$, is computed. Then, the first and third quartiles (i.e. $25^{\text{th}}$ and $75^{\text{th}}$ percentiles) of the wind speeds observed in this direction are used to obtain an estimation of the minimal and maximal wind speeds (respectively $V_{i\to x}^{min}$ and $V_{i\to x}^{max}$). Wind speeds at a 850 hPa pressure level (Figure 4.30a) are considered because this value corresponds to the altitude at which cumulus clouds usually form [62]. The minimal and maximal propagation times required for the information to travel from one site to another [2] are then obtained considering Equation 5.2, where $D_{x\leftrightarrow i}$ is the distance between the site of interest, $x$, and its neighbour $i$. Considering the forecast horizon $h$ and the possible lag $\tau_{lag}$, the relevant information is assumed to be contained within the range defined by Equation 5.3.

$$\begin{cases} T_{i\to x}^{min} = \frac{D_{x\leftrightarrow i}}{V_{i\to x}^{max}} \\ T_{i\to x}^{max} = \frac{D_{x\leftrightarrow i}}{V_{i\to x}^{min}} \end{cases} \tag{5.2}$$

$$T_{i\to x}^{min} \leq \tau_{lag} + h \leq T_{i\to x}^{max} \tag{5.3}$$

Forecasting performances obtained with this feature preselection process are depicted by the blue curve in Figure 5.3 (i.e. the *RF + SDU(t-9:t/physics-constrained)* model). We observe that the proposed preselection approach enhances both nRMSE and nMAE forecasting scores for horizons up to 2 hours ahead, and reduces computational time by around 40% in comparison with the model considering all the lags and close neighbours (i.e. model *RF + SDU(t-9:t)*). A slight degradation in the nRMSE scores is observed for greater horizons.

A similar investigation was conducted with the AR model. Contrary to the RF model, the physics-constrained preselection method did not contribute to improve forecasting accuracy but was retained on account of the Occam's razor and a reduced computational time.

#### 5.2.4.3 Feature importance

To get an insight into the inner workings of the forecasting model, we focus on the importance of the features attributed by the *RF + SDU(t-9:t)* model (Figure 5.4). As a reminder, the determination of feature importance is presented in Section 2.2.4.3. An explanation of the prevalence of some features can be found in the wind distribution analysis performed in Section 4.4.2.5, where we highlighted the presence of three main wind regimes. The Mistral is a dominant stream coming from the north. This wind does not play a significant role in the transport of clouds, unlike the south wind, which is associated with moisture-laden air. At higher layers of the atmosphere, western weather structure displacements associated with oceanic air flux play a predominant role in the transport of cloud structures. The

1. Angle between the direction of the two plants and that of the north.
2. We adopt a Lagrangian description of the cloud movement.

spatial distribution of PV plants is presented in Figure 2.5.

An analysis of feature importance reveals that the *RF + SDU(t-9:t)* model tends to give relative high importance to PV8 for the 15-min ahead horizon. This prevalence is in line with the distribution of the south wind, but is surprising given the distance between PV1 and PV8. Despite their distances, PV8-derived features contribute more to the forecast of PV1 production than its closest neighbour PV3. The graph shows that the ST information from PV7, PV5 and PV6 contributes little to forecasts performed at PV1 compared to other sites such as PV3. This is due to (1) the spatial location of the sites: PV5 and PV6 are located northward in the direction of the Mistral wind, while PV3 is westward, and (2) PV3 is closer than the two other sites. Observations made on PV7, which is the furthest site, have very little importance in the forecast accuracy of PV1 production for 15-min ahead, but feature importance slightly increases as the forecast horizon gets higher. This may be explained by the propagation speed of the ST information from one site to another. Similar observations are made with PV5 and PV6 features, which are located in the same direction as PV7. This indicates the robustness of the results.

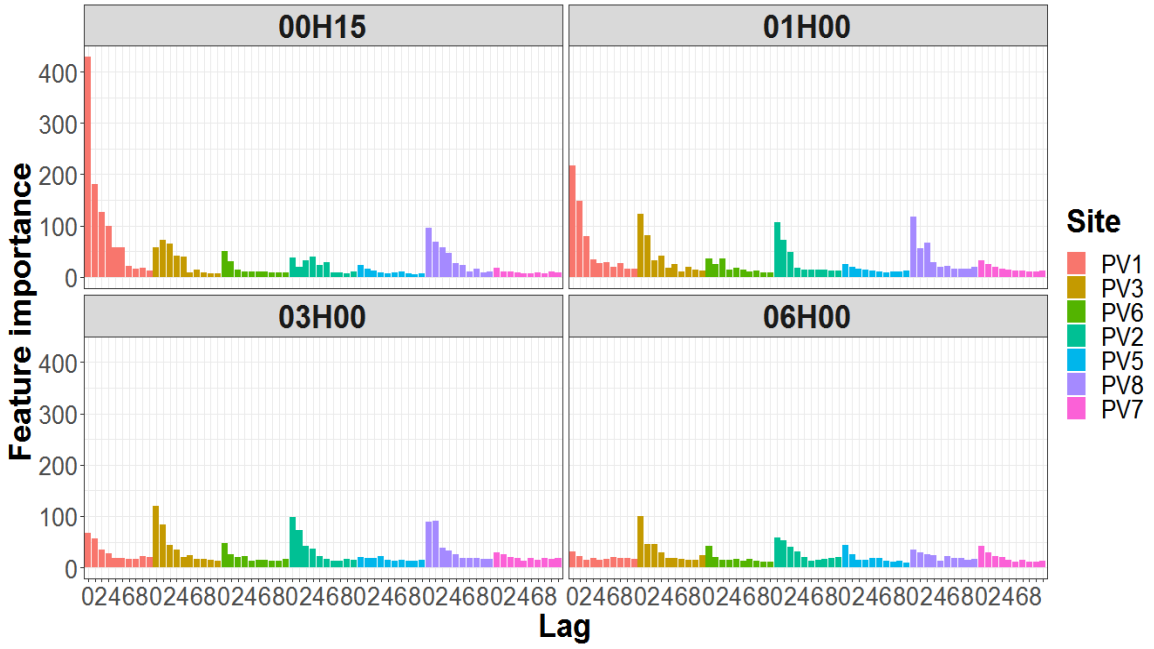



Figure 5.4 – Feature importance computed for the 15-min, 1-hour, 3-hour and 6-hour forecast horizons from the RF models considering PV1 and its closest neighbours (i.e. model *RF + SDU(lag 0:9)*). Colours represent the power units (ranked from the closest to the furthest) while the x axis stands for the temporal lag of the features.

It is interesting to note that for the 3-hour and 6-hour ahead horizons, the last observations at PV3 (i.e. observations associated with lag 0) have more importance than those from PV1. For the 15-min ahead horizon, we observe that the most informative lag at PV3 is the very first, while PV2 is characterised by a bi-modal distribution with two modes present at lag 0 and 3. These may illustrate the use of ST dependencies by the model. Nonetheless,

we observe that in general the feature importance decreases as the lag gets higher for the four forecast horizons to such an extent that the last observations are usually the most informative ones. This behaviour is quite surprising. As the ST information has a finite speed of propagation imposed by wind speed, we were expecting former lags to gain more importance for higher lead times. A hypothesis to explain this phenomenon is that it is hard for the fixed-coefficients regression model to identify ST dependencies between power units due to the wide range of weather dynamics induced by clouds directions and speeds. As a result, it focuses mainly on the last available observations to derive spatial dependencies.

#### 5.2.4.4 Additional restrictions on lagged observations

This motivates us to test additional restrictions on the generation of inputs lags to retain only the last observation, which is in general the most informative whatever the forecast horizon considered. This approach achieves the best forecasting performances in terms of nRMSE and nMAE for horizons lower than 120-min ahead (pink curve in Figure 5.3). Significant improvements are observed for the 15-min ahead horizons between the models *RF + SDU(t)* and *RF + SDU(t-9:t)* reaching 4% and 8% in terms of nRMSE and nMAE respectively. Nevertheless, a slight drop in the nRMSE scores is observed for higher horizons. This tends to support that the learning of the RF model is mainly based on the recognition of similar spatial patterns rather than on the extrapolation of cloud motion based on ST correlations. Henceforth, we only consider the RF fed with the last production level observed on neighbouring plants. The terminology *RF + SDU* implicitly refers to *RF + SDU(t)*.

### 5.2.5 Comparison of models' accuracy

In this section we compare the forecasting performances of the AR and RF models fed with temporal- or ST-based inputs (Figure 5.5).

The inclusion of distributed PV observations improves forecasting performances on the whole horizon spectrum in comparison with temporal-based approaches for both models under consideration. Yet, we observe that the ST information slightly degrades the bias of the AR models. Contrary to the persistence model, the considered approaches tend to over-forecast. We observe a slight degradation in the nRMSE score for the ST version of the AR models at 6-hour ahead horizon. In accordance with what has been said previously, this is thought to result from an overestimation of the spatial de-correlation distance performed in Section 5.2.3. Overall, the performance improvement due to ST information is higher in the case of RF for the three scores under consideration. For both AR and RF models, we observe that the improvement resulting from neighbouring observations reaches its peak at around the 1-hour ahead horizon, then progressively decreases to become negligible at the 6-hour ahead horizon. This is explained by the fact that for the very first forecast horizons, the most relevant source of information is provided by previous production measurements at the site location (due to the weather persistence), while for the highest forecast horizons,

no relevant information is extracted from the sensor network because of the temporal decorrelation distance and the chaotic nature of weather.

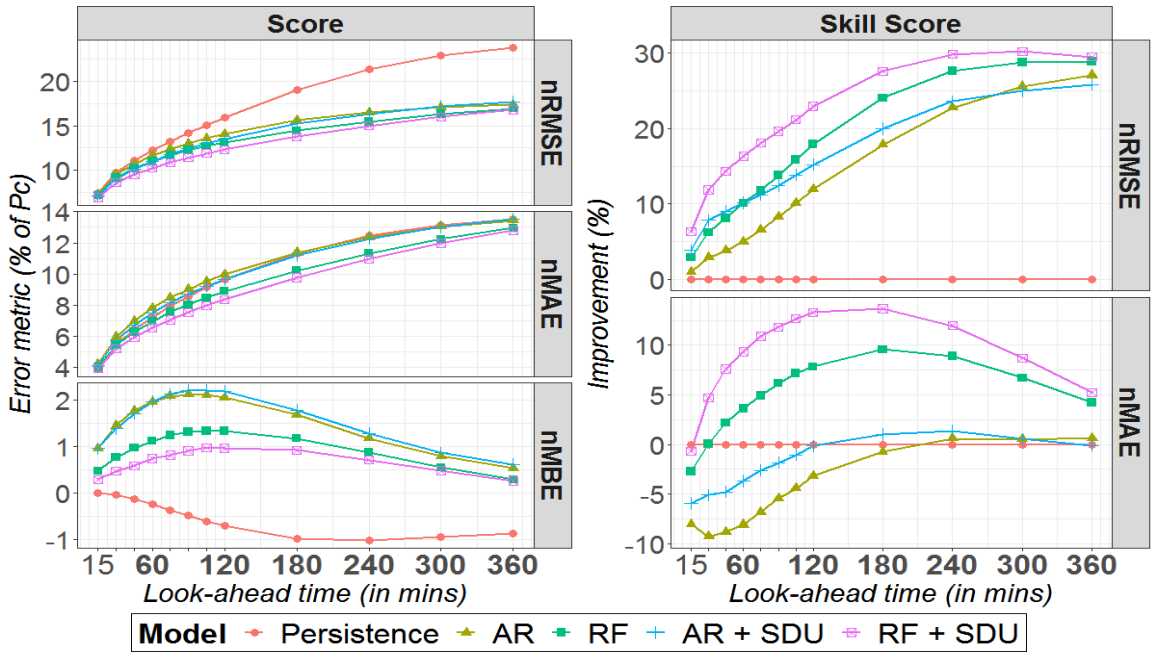



Figure 5.5 – Forecasting performances of the AR and RF models fed with observations at the sites of interest and their ST versions.

The *AR + SDU* model improves the nRMSE and nMAE scores up to 5% and 4% respectively, while for the same metrics the *RF + SDU* model reaches improvements of 7% and 6%. Bessa et al. in [83] compare performances of a VAR model considering Distribution Transformer Controller (DTC) measurements and a Vector Auto-Regressive with eXogenous inputs (VARX) model, which also integrates sensor observations. Figure 5 from the aforementioned article shows that the global nRMSE improvement of VAR over AR varies between 4.2% and 2%, while improvements due to the VARX framework range from 9% and 5.7%. These findings corroborate the performances shown in Figure 5.5 and suggest that a denser source of information benefits forecast accuracy.

This section underscores the two main flaws inherent to the PV production network under study: its low density and its north-south orientation which prevents observations of western winds. To fill this gap, information from satellites is considered.

## 5.3 Irradiance satellite-based information

### 5.3.1 State-of-the-art

The literature employs several approaches to deal with satellite-based information. One of the most widespread methods consists in extrapolating cloud displacement using motion extraction techniques developed in the image processing field. Thus, a block-matching method applied to two successive images makes it possible to identify positions of similar

cloud structures, and then to derive displacement vectors. Cloud Motion Vector (CMV) are then used to translate the most recent map by assuming that cloud structure remains unchanged over time [247]. CMV-based methods reveal interesting forecasting performances up to 2 hours ahead. Forecasting performances can be extended to further horizons by considering wind velocity computed by Numerical Weather Predictions (NWPs) models as displacement vectors. The main drawback of CMV approaches lies in their ineffectiveness in the case of local cloud formations [248]. More recent studies resort to statistical or deep learning approaches. For instance, [249] proposes a straightforward modelling chain, where a satellite image is flattened and fed into an SVM model which provides a forecast of PV production. In [250], the authors propose a Deep Neural Networks (DNN) architecture, which extracts relevant features from three consecutive satellite images (via Convolutional Neural Networks (CNN)), which are then combined with meteorological data and fed into an ANN to derive irradiance forecasts.

The use of satellite-based information drastically increases the number of features involved in the forecasting process, which at the same time increases the computational burden. Here, the number of features of a dataset is referred to as its dimensionality. As the dimensionality of a dataset extends, it becomes more and more difficult to derive accurate forecasts from the dataset; this is called the curse of dimensionality. A related notion named *statistical curse of dimensionality* implies that to obtain statistically reliable results, the sample size grows exponentially with the data dimension. For these reasons, a reduction of the dimensionality of satellite-based information is needed while preserving relevant information.

**Research Gap - *Dimensionality issue***

Contrary to the spatial inflexibility inherent to PV networks, satellite-based observations offer the possibility of covering the whole vicinity of the site location, and much more. This raises new issues; how can we efficiently capture the ST dynamics, while reducing the dimensionality and avoiding the risk of overfitting?

### 5.3.2 Dimensionality reduction

The very first step to reduce the dimensionality of the satellite-derived maps is to limit their radius. However, this approach is not sufficient: considering a radius of 100 km implies injecting of 900 variables in the forecasting models (without counting potential features lags).

In this section, three preprocessing options are investigated with the forecasting architecture defined in Figure 5.6. The first one aims at selecting a subset of $N$ features considered as relevant according to some criteria. The second approach projects the high-dimensional data to a space with fewer dimensions. The last option proposes a dedicated model, which derives forecasts at the site location from previous satellite observations.

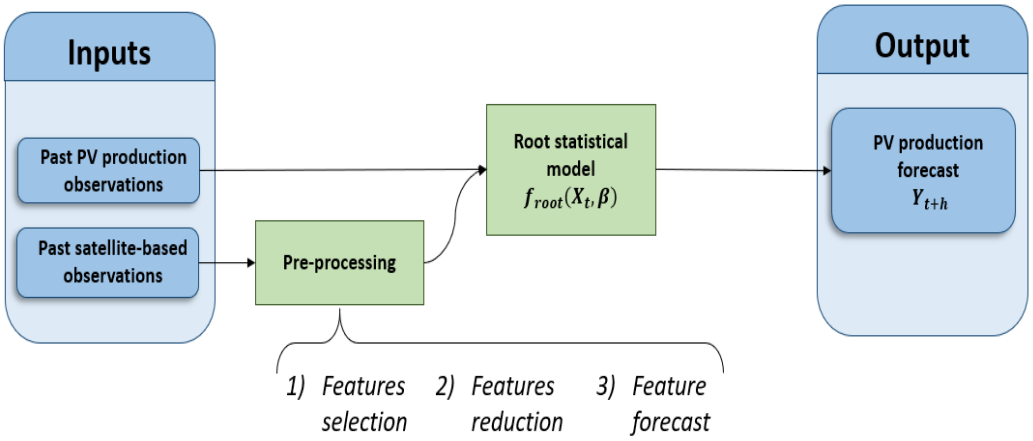



Figure 5.6 – Workflow of the forecasting architecture studied in this section.

#### 5.3.2.1 Feature selection

To cope with this dimensional burden, it is common practice to resort to a straightforward approach: we can consider a set of well-chosen pixels fed to the forecasting model. This selection step is performed on the training set in order to provide a subspace composed of $N^{SDSI}$ satellite pixels. Usually, an Maximal Relevance Feature Selection (MRFS) selection scheme is used [55, 67]. This scheme consists in selecting a user-defined number, $N^{SDSI}$, of features that have the highest correlation with the target variable. In other words, a correlation scores analysis performed between satellite-derived information and $h$ time-led PV production is used to select pixels that have the highest scores depending on the forecast horizon, $h$. The Pearson correlation criterion is often used in feature selection processes [55, 236], while [67, 71] consider the MI criterion for its ability to identify nonlinear relationships. Here, we focus on the MI criterion and on an alternative feature selection scheme, namely the minimal-Redundancy-Maximal-Relevance (mRMR) scheme. We do not include results obtained with Pearson-based selection, because they are, overall, inferior to those obtained with MI-based selection [71]. The focus is rather on the selection scheme rather than on the correlation criterion. Figure 5.7 provides a graphic summary of the considered options.

**5.3.2.1.1 Mutual information criterion** An MRFS selection scheme is usually implemented: we uses a score to measure individually the dependence between the explanatory features and the target variable, then the $N^{SDSI}$ features that have the highest relevance are kept.

Carriere in [67] proposes to use the MI criterion instead of the Pearson correlation score because of its capacity to assess nonlinear relationships. This criterion measures the dependence between two random variables $A$ and $B$, and more precisely the reduction of uncertainties regarding one variable when the other one is known. The higher the MI, the higher the reduction in uncertainties, while a zero MI indicates features independence. For

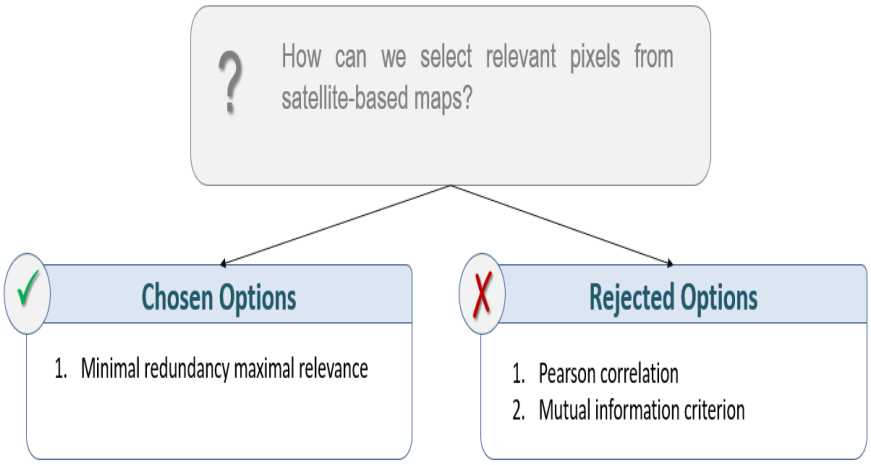



Figure 5.7 – Considered feature selection methods.

two discrete random variables, the MI is computed as follows:

$$MI(A,B) = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p_{A,B}(a,b)\log\left(\frac{p_{A,B}(a,b)}{p_A(a)p_B(b)}\right). \tag{5.4}$$

| | |
|---|---|
| $(A,B)$ | A pair of discrete random variables with values over the space $\mathcal{A}\times\mathcal{B}$, |
| $p_{A,B}$ | Joint probability distribution of $A$ and $B$, |
| $p_A$ | Marginal probability distribution of $A$. |

Figure 5.8 represents the MI score computed for each pixel from the SDSI dataset with the PV production observations for several forecast horizons. It reveals that the correlation area is highly influenced by the topography of the region (i.e. the funnel-shaped region is due to the Rhone Valley, Figure 5.9). Two main observations can be made. First, this approach tends to select westward points for 1-hour and 2-hour ahead time-steps. This finding supports the analysis performed in Section 4.4.2.5, which suggests a predominance of ST structures coming from the west. Second, the selected points tend to cluster. This observation is all the more valid as the forecast horizon is low.

The main limitations to this approach become apparent when it comes to forecasting PV production with cloud motions that do not result from the prevailing wind direction. One option could be to increase the number of selected points, but it would hardly improve forecasting performances due to the pixel aggregation phenomenon. To illustrate this statement, we can say that the MI-based selection process makes the forecasting model blind in some spatial directions while providing satellite pixels carrying redundant information.

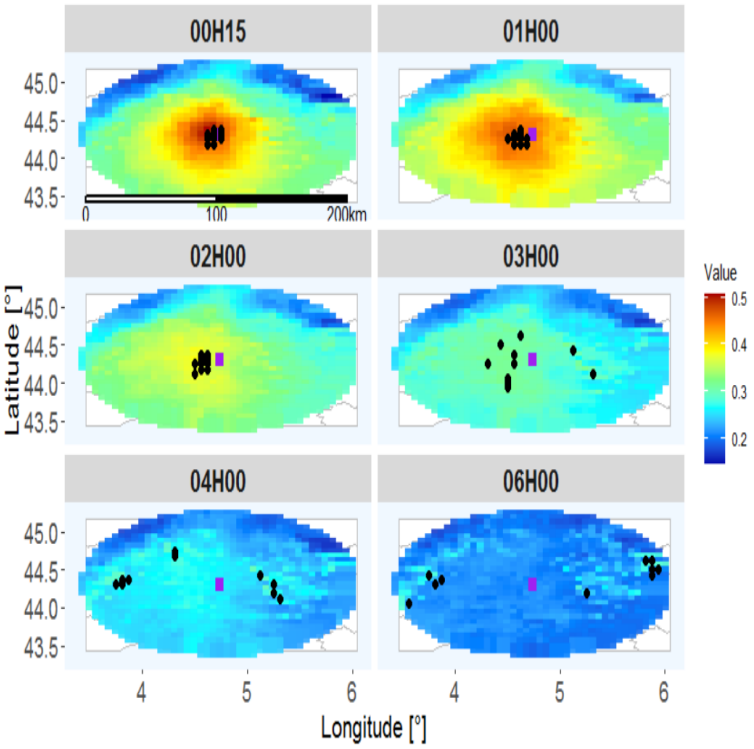



Figure 5.8 – Position of the 10 satellite pixels that have the highest MI score (in black) with PV production observations for PV1 (in purple) for different forecast horizons. The background represents the annual inter-correlation map for all grid points.

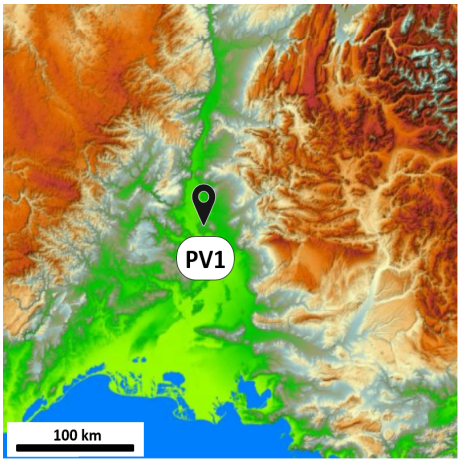



Figure 5.9 – Topography of the Rhone Valley. Relief map generated from https://www.geoportail.gouv.fr.

**5.3.2.1.2 Minimal-redundance maximal-relevance** To address the issue of redundancy among selected features, the mRMR incremental selection framework [251] is implemented. An mRMR scheme is usually applied with gene expression data. This is an incremental selection method that aims at finding a subspace of $N^{SDSI}$ features that minimise the MI criterion between selected features while maximising the MI criterion between each selected feature and the target situation. Here, we consider a forward selection scheme which intends to incrementally select $N^{SDSI}$ features from a features pool $S_M$ containing $M$ variables. First, the algorithm selects the feature $s_1$ that has the highest MI with the target

variable. Then, a second feature, $s_2$ is selected from the set $S_M - s_1$ in such a way that the redundancy within $Z_2 = \{s_1, s_2\}$ is minimised while the MI score between $s_2$ and the target is maximal. This incremental feature selection process is repeated until the number of iterations $N^{SDSI}$ is reached.

Figure 5.10 displays the ten satellite pixels selected by the mRMR approach. Unlike the MI selection, this procedure tends to select pixels in every direction. We can note that for short-term forecast horizons, the selected points are no longer aggregated near the farm but are located at some distance, which can be beneficial in an ST forecast context involving several weather dynamics.

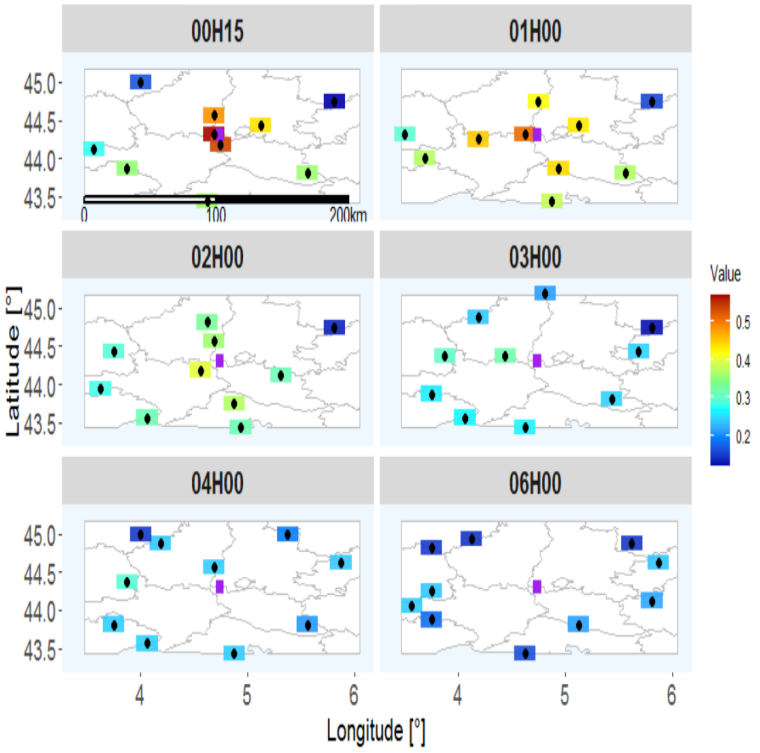



Figure 5.10 – Position of the 10 satellite pixels (in black) selected via the mRMR selection scheme for PV1 (in purple).

In addition, this graph shows one limitation in this approach. To comply with the user-defined number of pixels, some low-informative pixels are selected; for instance, the two pixels with an MI lower than 0.2 for the first time-step. As observed in the previous section, this may be detrimental to forecast accuracy.

**5.3.2.1.3 Forecasting performances** First, it is necessary to determine the optimal number of features (or pixels) to include in the ST model (1) to obtain the best forecasting performances, and (2) to avoid over-fitting. A sensitivity analysis performed on the number of SDSI features ($N^{SDSI} \in [\![5, 50]\!]$) reveals that MI-based and mRMR-based selections perform better when considering the first 10 features with the highest dependence scores. Beyond these values, the increase in features does not improve forecasting performances.

Similarly to the previous section, a cross-validation approach highlights that better forecasting performances are reached when considering lagged SDSI-features with the AR model, while an opposite trend is observed with the RF model. In this latter case, only the last observations of SDSI-based features are considered. It turns out that the AR model fed with SDSI features is less parsimonious than the RF-based approach. This phenomenon may be attributed to the nonlinear structure of RF. But in any case, the forecaster needs to remain attentive because the embedded feature selection process of RF may reach its limits.

Figure 5.11 represents forecasting performances of the AR and RF models fed with the sets of SDSI features provided by each feature selection process. First, we observe that the feature selection approaches have very little influence for forecast horizons lower than 1-hour ahead. Nevertheless, the mRMR-based selection process exhibits higher skills for higher horizons both with the AR and RF models. In the case of the RF model the nRMSE score enhancement can reach around 4%. Such results have already been presented in [71] but differ slightly in the performances gain obtained with the mRMR method compared with MI-based selection. The main differences between the results presented in this section and the former publication lie in the clear-sky normalisation process (addition of the physics-based modelling of irradiance), and the correction of abnormal production observations.

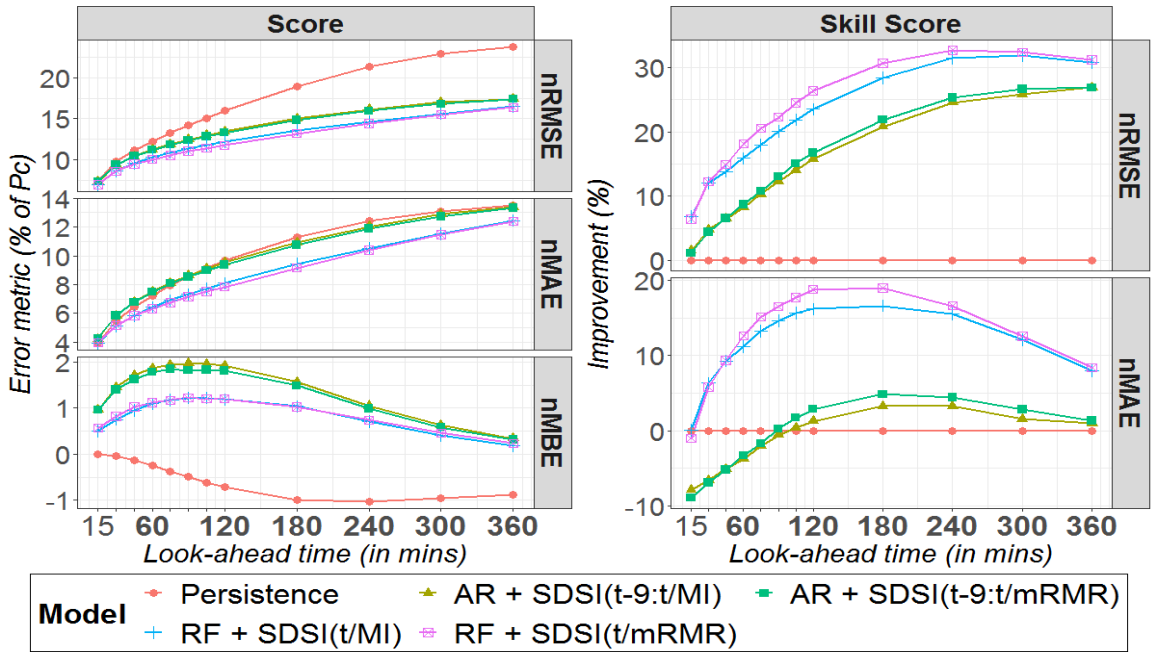



Figure 5.11 – Influence of the SDSI feature selection processes over the forecasting performance of the AR and RF models. The number of lags considered for the SDSI features as well as the feature selection approach are given in brackets.

**Research Answer - *Dimensionality issue (1/2)***

Within the scope of feature selection, the mRMR framework improves the forecasting accuracy of both linear and nonlinear models compared to other methods based on the Pearson-based correlation or the MI score, which are traditionally found in the literature.

**5.3.2.1.4 Selected features** Let us focus on the importance given to each SDSI feature by the *RF + SDSI(t/mRMR)* model (Figure 5.12). For the first time-step, the embedded feature selection approach mainly selects observations from the two closest pixels located westward and southward. In this case, the farthest features are irrelevant inasmuch as the main information is still carried by previous observations performed at the site's location. As the forecast horizon gets higher, the algorithm tends to widen its spectrum of selected features. We observe that for the 1-hour and 3-hour ahead horizons, more importance is also given to westward located features. On the whole, little importance is given to pixels located in the north, which is in line with the wind regime analysis performed in Section 4.4.2.5.

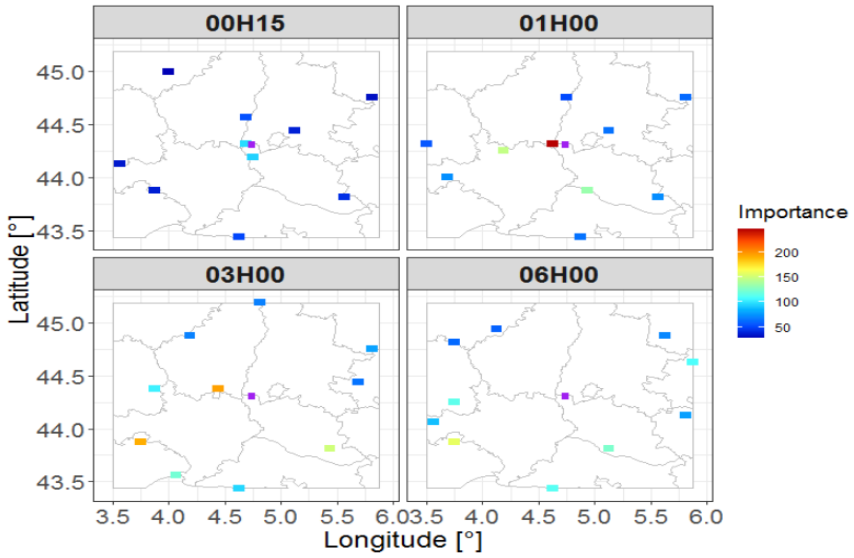



Figure 5.12 – RF-based importance of features selected with the mRMR framework.

### 5.3.2.2 Feature reduction

After selecting a set of relevant features that carry the meaningful information contained in the SDSI maps, in this section, we investigate the transformation of data from high-dimensional spaces to low-dimensional representations, by means of a Principal Component Analysis (PCA) [252].

PCA is an orthogonal linear transformation that transfers the data into a new coordinate system. Let us consider a matrix where features are stored in columns (thus, rows represent observations). Each feature of the dataset is considered as an individual dimension of a

$N^{PCA}$ feature space. First, the data are centred so as to find the coordinate system origin. Depending on the nature of the features, the latter can also be reduced (e.g. if features are measured on different scales). Then, the main idea behind PCA consists in finding an orthogonal basis constituted by $N^{PCA}$ dimensions, also called Principal Components (PCs), in such a way that they capture the maximum data variation (i.e. these dimensions are oriented in the direction of the largest variance in the dataset). Thus, keeping only the $N^{PCA*}$ PCs, which explains most of the variance, reduces the dataset dimension while minimising the information loss.

To obtain the PCs, first the co-variance matrix (if the features are centred) or respectively the correlation matrix (if the features are standardised) is computed. This matrix describes the dispersion of the measured features. Then, the eigenvectors and the corresponding eigenvalues of the co-variance/correlation matrix are computed through a Singular-Value Decomposition (SVD). The resulting eigenvectors, which are unit orthogonal vectors, correspond to the PCs, while the eigenvalues quantify the importance of the PCs in terms of explained variances. Thus, sorting the eigenvalues into a descending order allows us to assess the importance of each PC. PCs accounting for a cumulative explained variance of 90% are usually kept, while the others are discarded. Lastly, the original dataset is projected into this new dimension-reduced space.

Considering 100-km radius SDSI maps and a 90% threshold on the cumulative explained variance, 13 PCs are retained out of the 903 initial features (in this context, a feature represents a time series derived from the value observed at a specific pixel). This roughly represents a dimension reduction of 99%. Figure 5.13 represents the explained variance of the 20 first PCs and their cumulative sum.

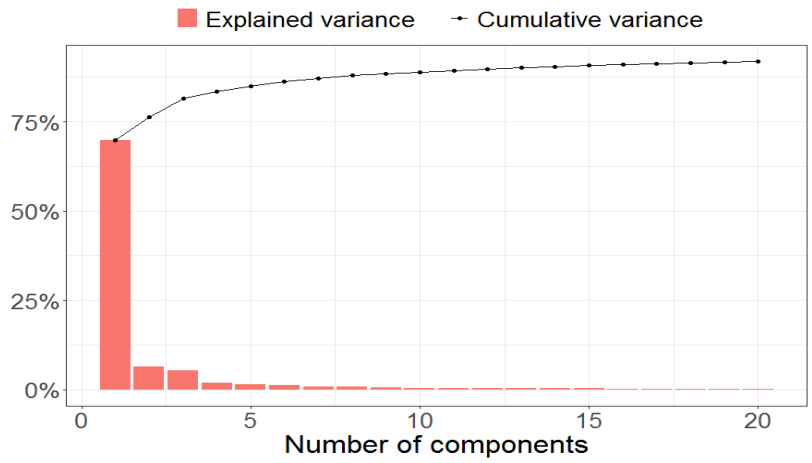



Figure 5.13 – Explained variance of each PC and their cumulative sum. Only the 20 first PCs are represented.

The resulting $N^{PCA*}$ projected features are injected into the root forecasting models as additional explanatory features.

#### 5.3.2.3 Feature forecast

Over the last few years, we observe a growing use of DNN algorithms in the renewable energy forecasting field. Deep learning refers to a network composed of multiple stacked layers (an input layer, an output layer and at least one hidden layer in between) able to extract high-level features from inputs, which reduces the need for feature engineering. Deep structures tend to achieve better forecasting accuracy compared to single layer models for their complex nonlinear mapping capabilities. In addition to their proven interest for datasets containing images or problems related to classification, deep neural networks tend to dominate forecasting competition together with gradient-boosted decision trees [253].

The introduction of Convolutional Neural Networks (CNN) [254] opened the door to the development of computer vision techniques. The main advantage of a CNN over an ANN is its ability to work with 2D data, which preserves spatial patterns contained in the images. Successive convolutional layers composed of convolutional filters allow the identification of main features ranging from low-level features (e.g. lines) to more abstract features (e.g. shapes or objects) in the higher layers. CNN has already been successively applied to the solar power domain to extract relevant information from satellite images [250] or to forecast production [255]. To capture patterns in sequences of data, such as time series data, Long Short-Term Memory (LSTM) networks are a good option. LSTM are an extension of Recurrent Neural Networks (RNN) able to retain information for long periods of time thanks to their memory cell, which acts as an accumulator of the state of information. Since CNN and LSTM are well-known methodologies, theoretical explanations of these models are omitted but interested readers may refer to [256, 257].

To benefit from the feature extraction ability of the CNN and the memory capacity of the LSTM network, a hybrid modelling has been developed. The CNN-LSTM [257] is composed of a sequence of several convolutional layers, and a flattening layer used to provide 1D data to the LSTM layer. The authors prove that the proposed model outperforms both CNN and LSTM architectures for day-ahead forecasts. For shorter horizons, namely 1-hour ahead, [258] shows that CNN-LSTM network outperforms CNN but that both models are less accurate than the Auto Regressive Moving Average (ARMA) model. The major drawback of this approach in handling ST information is that it is necessary to unfold the data to 1D vectors to enable LSTM network processing. This step loses spatial information. The ConvLSTM [256] on the other hand, is a specific layer dedicated to ST sequence forecasting problems. In short, the ConvLSTM layer is similar to the LSTM layer except that matrix multiplications are replaced by convolution operations, which allows us to keep the spatial dimension of data. The authors highlight that the ConvLSTM network captures ST correlations better and consistently outperforms LSTM. For a comprehensive description of the mathematical formulation of CNN-LSTM and ConvLSTM, interested readers may refer to [259]. In [259], a comparison of the forecasting performance of both hybrid models is carried out with PV production data. The results show that for the one-day-ahead time

horizon, the ConvLSTM-based architecture provides better forecasting performances. It is worth mentioning that in this work both architectures slightly differ: the ConvLSTM-based architecture possesses only one ConvLSTM layer, while the CNN-LSTM network has three convolutional layers.

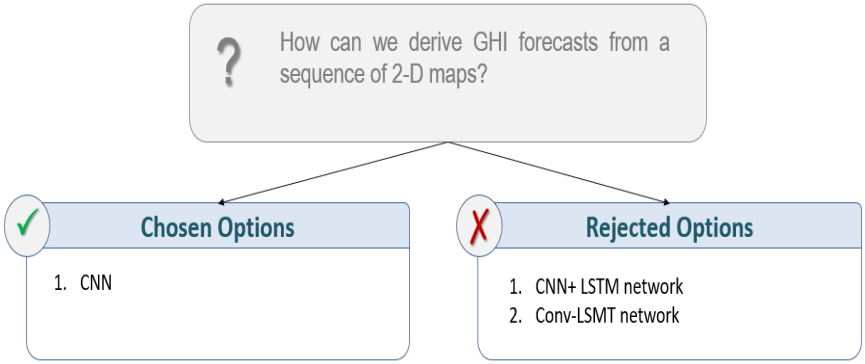



Figure 5.14 – Set of options investigated to extract relevant information from SDSI with DNN-based models.

The CNN and Conv-LSTM architectures have been investigated (Figure 5.14) to derive Global Horizontal Irradiance (GHI) forecasts at the plant location from SDSI maps. The main idea is to forecast GHI at the PV plant location pixel using a *ST cube* representing the temporal sequence of last observed 2D maps. This *ST cube* is then fed into a DNN composed of $n$ layers. The design of this network is inspired from what can be found in the literature (e.g. [250]) and from a trial and error process. Each layer is first composed of a convolutional/Conv-LSTM layer, directly followed by a batch normalisation layer. The latter is known to speed up training and to improve performances by re-centring and re-scaling inputs [260]. Then, a pooling layer is traditionally used, which downsamples the feature map generated by a convolution layer by aggregating the features present in different regions of the map (e.g. the maximum element from a region is selected). Despite observing a fitting time reduction, empirical testing shows that better forecasting performances are achieved by replacing this layer with a dropout layer. The dropout layer is a regularisation technique used to improve training performances, which has been proven effective in reducing over-fitting [260] by preventing a fraction (here 20%) of neurons from training at each iteration. The last stage of a CNN-based architecture is composed of the flattening and dense (also called fully-connected) layers. The flattening layer converts the data into a 1-dimensional array, which is then injected into the dense layer. This is a layer that connects every neuron in one layer to every neuron in another layer. The dense layer compiles the data extracted by previous layers to form the prediction, per se. It is in this last layer that the regression work is performed. The DNN architectures are illustrated in Figure 5.15. These layers utilise the Rectified Linear Unit (ReLU) activation function. The network is fitted by a state-of-the-art optimiser, namely the Adam optimisation algorithm [261], while the Mean Square Error (MSE) of irradiance is defined as the loss function.

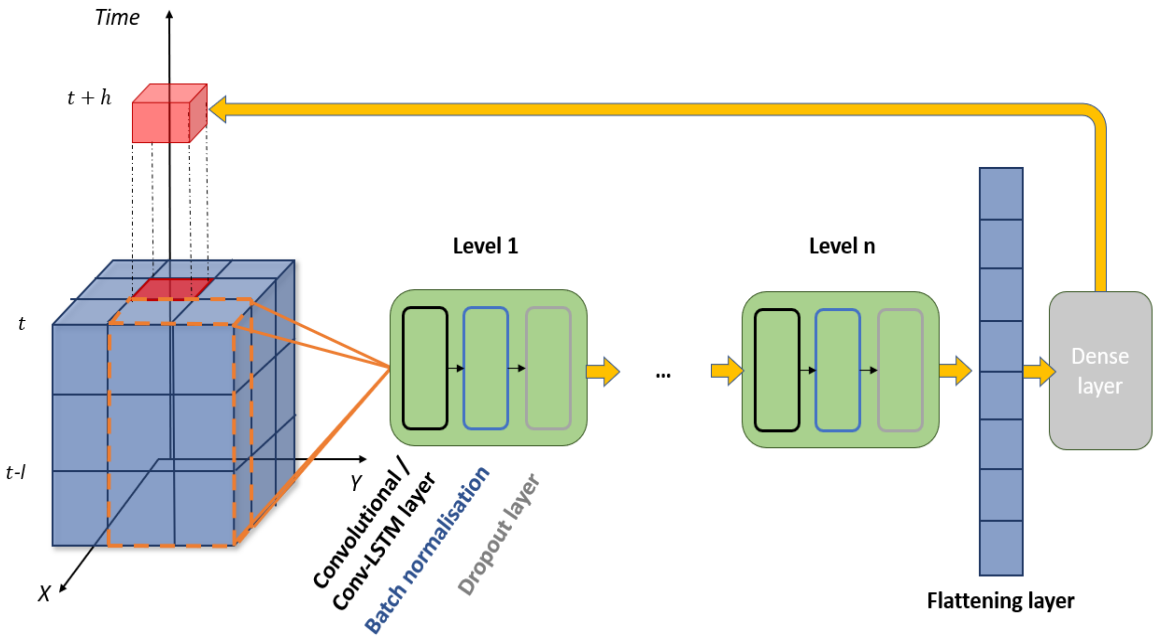



Figure 5.15 – Topological structure of the CNN/Conv-LSTM algorithms integrated with the objective of predicting solar irradiance at the site position (red area) at time $t + h$. The ST cube of past SDSI observations is used to train the CNN/Conv-LSTM models.

In this work, 4 and 3 levels are considered respectively for the CNN- and the Conv-LSTM-based architectures. The high computational cost associated with this last model is a significant issue: it takes up to 4 days to fit the Conv-LSTM-based model for a specific site and the 12 forecast horizons under study. To simplify the modelling process and to reduce the computational training time, Transfer Learning (TL) [262] is implemented. This Machine Learning (ML) technique receives considerable attention in the literature because it solves the problem of limited data or insufficient computer resources, and reduces fitting time. The main idea behind this method is to pre-train a DNN for a specific task, and then to transfer the knowledge learnt to another related field by performing fine-tuning of the deepest network's layers, while the first layers, acting as a general feature extractor, remain untouched. For instance, a possible application of TL could be the re-training of a dog classifier into a cat classifier. Here, a Conv-LSTM network is fitted on a specific plant, and then TL is used to fine-tune the model for other locations. We choose to fit a model for each site to integrate potential local weather dependencies.

Subsequently, models are trained over years ranging from 2010 up to 2014 to derive GHI forecasts at PV plant locations for 2015 and 2016. The latter are injected in the AR and RF models together with past PV production observations to forecast PV production for 2016. Figure 5.16 shows that the RF model fed with the GHI forecast produced with the CNN slightly outperforms its counterpart for horizons greater than 1-hour ahead both in terms of nRMSE and nMAE, but performances are analogue for shorter horizons. Lastly, CNN-based forecasts are retained for the lower computational cost induced by the algorithm.

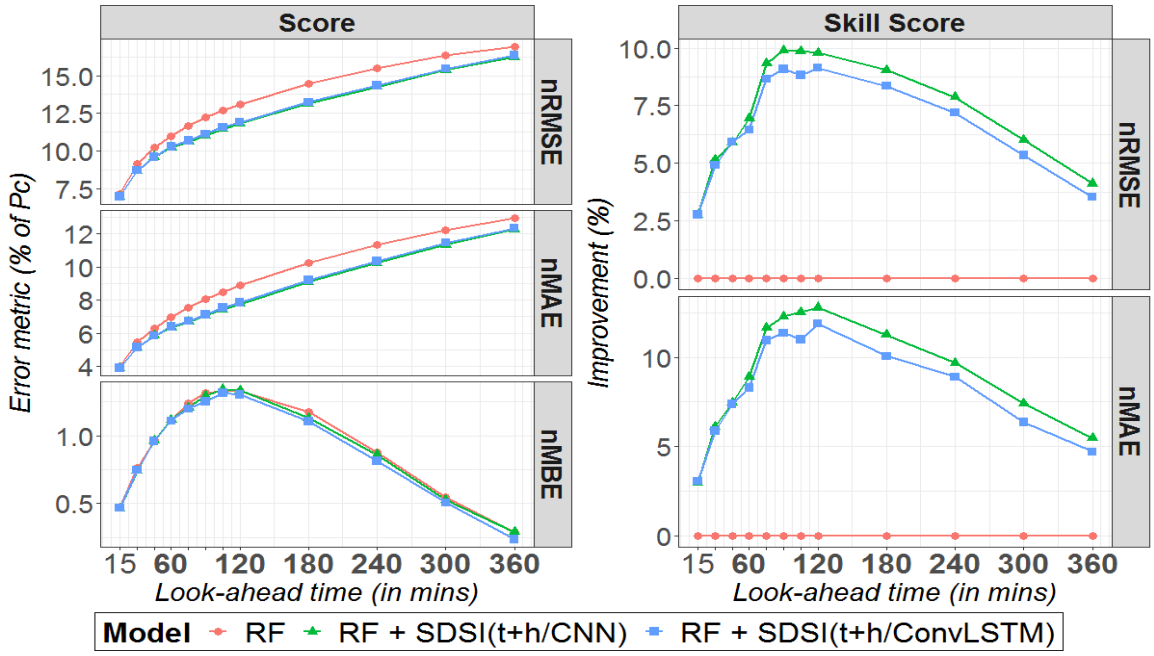



Figure 5.16 – Forecasting scores obtained with the RF model fed either with irradiance predictions issued by the CNN or the Conv-LSTM models.

### 5.3.3 Comparison of the different approaches

Outputs obtained from the features extraction and reduction models are now injected into the linear and nonlinear regression models. In addition, to get an insight of the full potential of SDSI observations, a theoretical variable standing for the forecast irradiance, which would be obtained with a perfect forecasting model from SDSI maps, is created. This new feature is simply obtained by considering SDSI observations at time $t + h$ rather than at a time prior or equal to $t$ and is, thereafter, denoted as a Satellite-based Perfect Forecast (SbPF).

Figure 5.17 and 5.18 gather respectively all the forecasting performances of the AR and RF models fed with the different outputs extracted from SDSI maps. To help the reader understand the nature of the inputs used in the regression models, the latter is explicitly stated in the legend ($t - 9 : t$ denotes that past observations from lag $t - 9$ to the current observations used, $t$ refers to the last observational data, while $t+h$ means forecast features).

First, the AR-based approach provides the widest range of observed performances depending on the feature processing method used (Figure 5.17). Forecasting performances are poorest when considering the feature selection algorithm, while the best scores are reached with forecasts issued by the CNN. In other words, complex preprocessing steps enable the extraction of relevant and easily assimilable information, while reducing the computational burden upon the AR model, which is observed in the form of a fitting time reduction. The relatively poor performances achieved by the PCA-based forecasts, and to a larger extent those obtained with the mRMR-based forecasts, highlight that cloud dynamics contained in SDSI maps are too complex to be modelled with a linear model. In case of the use

of CNN-based forecasts, we observe improvements due to the ST information that reach around 18% and 22% in terms of nRMSE and nMAE respectively.

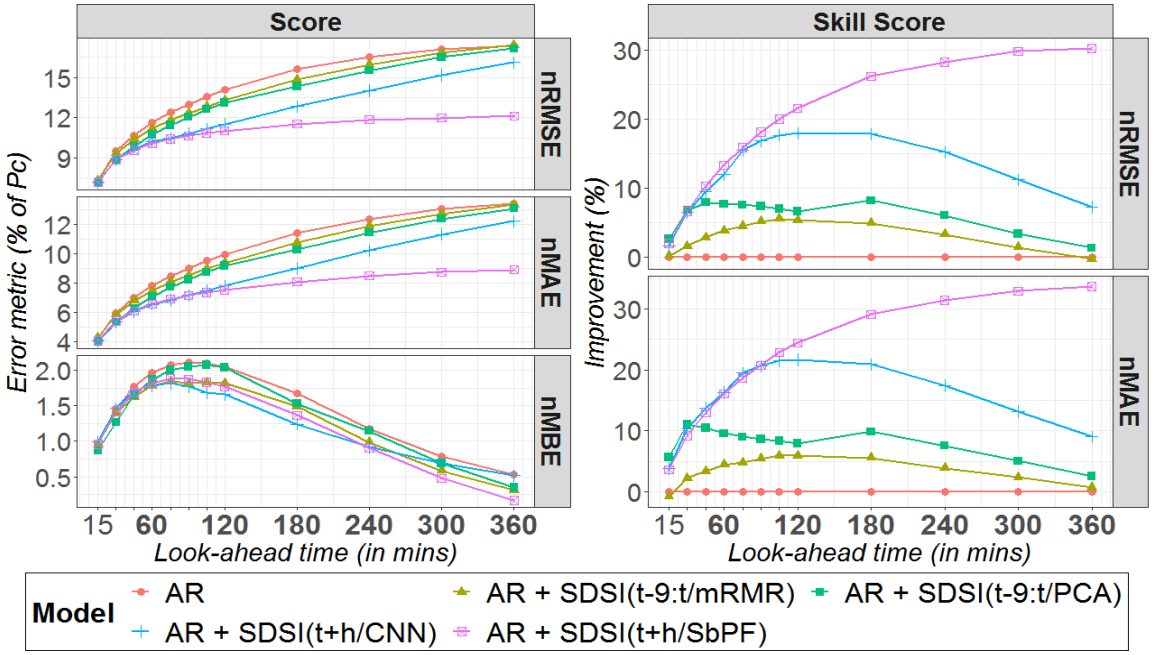



Figure 5.17 – Forecasting scores obtained with an AR model fed with past PV production observations and outputs from dimension reduction methods applied with SDSI. SbPF represents theoretical perfect irradiance forecasts at the site location from SDSI maps.

Regarding the use of the RF model, we observe in Figure 5.18 that the feature selection and reduction approaches have very little influence on the nRMSE and nMAE scores for all considered horizons. Yet, a closer look reveals that for horizons greater than 1-hour ahead, the CNN-derived production forecasts have a higher nMAE than the mRMR-based forecasts of around 1%. The use of PCA-based information tends to degrade the model bias.

As the nRMSE and nMAE differences are very low between the considered models, we implement the Diebold-Mariano (DM) test [99] (defined in Section 2.3.3) to judge the statistical significance of the differences. This test compares the predictive accuracy of two forecast models. Figure 5.19 highlights that independently from the horizons, all of the points (except one) are either outside the significance level of 5% or on the verge of being rejected. This signifies that the difference between the forecasts delivered by the three approaches is significant and that the information provided by the dimensionality reduction methods is not the same. This is confirmed by a deeper analysis that simultaneously combines outputs from the three proposed dimensionality reduction methods. This approach improves the nRMSE and nMAE scores by around 2 to 3% compared to the best performances obtained with the RF model fed with outputs from a unique dimensionality reduction method. However, this approach is not investigated in greater detail because the combination of the three reduction methods creates a huge computation effort for forecasters.

A comparison performed in Table 5.1 with a similar approach to the one used in this

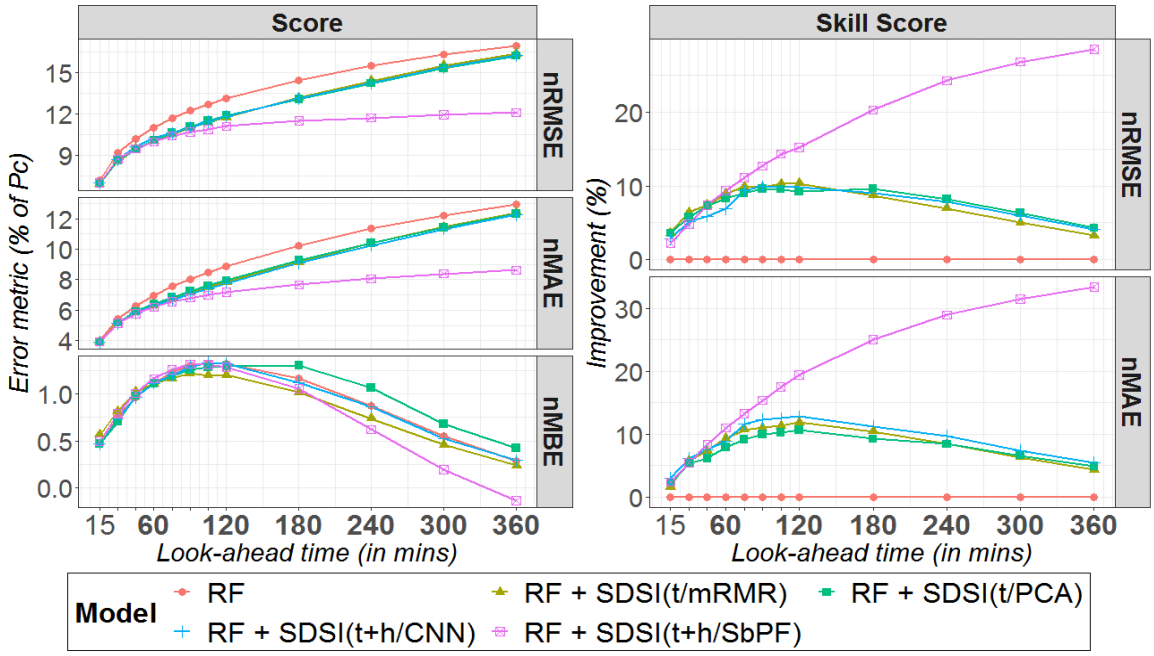



Figure 5.18 – Forecasting scores obtained with an RF model fed with past PV production observations and outputs from dimension reduction methods applied with SDSI. SbPF represents theoretical perfect irradiance forecasts at the site location from SDSI maps.

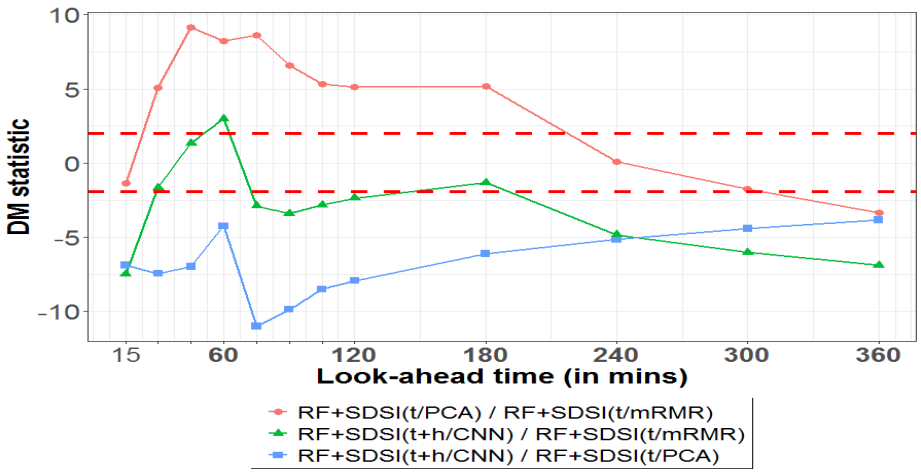



Figure 5.19 – DM test (defined in Section 2.3.3) between the three SDSI feature selection and reduction frameworks studied for different forecast horizons. The red dotted lines stand for the borders delimiting the validation and rejection of the null hypothesis.

section [3] suggests that the RF model is able to extract more relevant information from SDSI than an ANN used for horizons lower than 4 hours ahead. Unfortunately, based on the score used by the authors, we cannot develop the comparison of forecast accuracy any further due to the different nature of the quantities involved. In addition, it is difficult to draw definitive findings given the different climates under consideration.

Figure 5.17 proves that for short-term horizons (i.e. up to 105 mins) the coupling of the

3. In the sense that it is based on a nonlinear model fed with SDSI information and similar inputs except that the article focuses on irradiance forecasting.

| **Study** | [55] (Table 5) | Current study |
|---|---|---|
| **Location** | Gran Canaria Island | France |
| **Model compared** | *NN+SDSI / NN* | *RF+SDSI(t/mRMR) / RF* |
| **1-hour** | 4.99 | 9.32 |
| **2-hour** | 6.87 | 11.88 |
| **3-hour** | 7.75 | 10.42 |
| **4-hour** | 6.53 | 8.46 |
| **5-hour** | 8.16 | 6.30 |
| **6-hour** | 5.33 | 4.35 |

Table 5.1 – Comparison of nRMSE skill scores obtained with forecasts based on ST information w.r.t. predictions issued by temporal data from 1-hour to 6-hour ahead forecast horizons for different studies. Information within brackets represents the table containing values in the article. Authors in [55] forecast irradiance quantities.

AR model and DNN is able to extract all of the relevant information contained in SDSI maps. Indeed, *AR + SDSI(t+h/CNN)* and *AR + SDSI(t+h/SbPF)* models lead to similar skill scores on this range of horizons. Contrary to the RF model, the linear model used is not able to derive accurate, very short-term forecasts from the last observations (obtained through the mRMR pixels selection process or PCA performed on SDSI maps). This highlights that linear modelling is not suitable to account for the complex atmospheric phenomena occurring. In this case, the astutely fed RF offers an appealing alternative to the high computational effort involved in modelling ST dependencies through DNN architectures.

**Research Answer - *Dimensionality issue (2/2)***

In the case of RF, the choice of the dimension-reducing method has little influence over forecasting performances - compared to the AR model. The mRMR framework is recommended with this type of regression tool on the grounds of its computational efficiency. The AR model, due to its structure, is not able to extract the full ST information from PCA- or mRMR-based features. Thus, a nonlinear preprocessing of the SDSI maps is necessary. It is true that using the DNN introduced in the previous section to directly forecast PV production instead of GHI would have been easier, but it would have wasted the interpretability advantage of AR model.

At this point, it is insightful to compare the best integration strategies obtained with the AR and RF models. Figure 5.20 shows that the *AR + SDSI(t+h/CNN)* model outperforms the *RF + SDSI(t/mRMR)* model for forecast horizons higher than 90 minutes ahead in terms of nRMSE and nMAE. On the contrary, for shorter lead times, the *RF + SDSI(t/mRMR)* model prevails. This model also leads to the best bias.

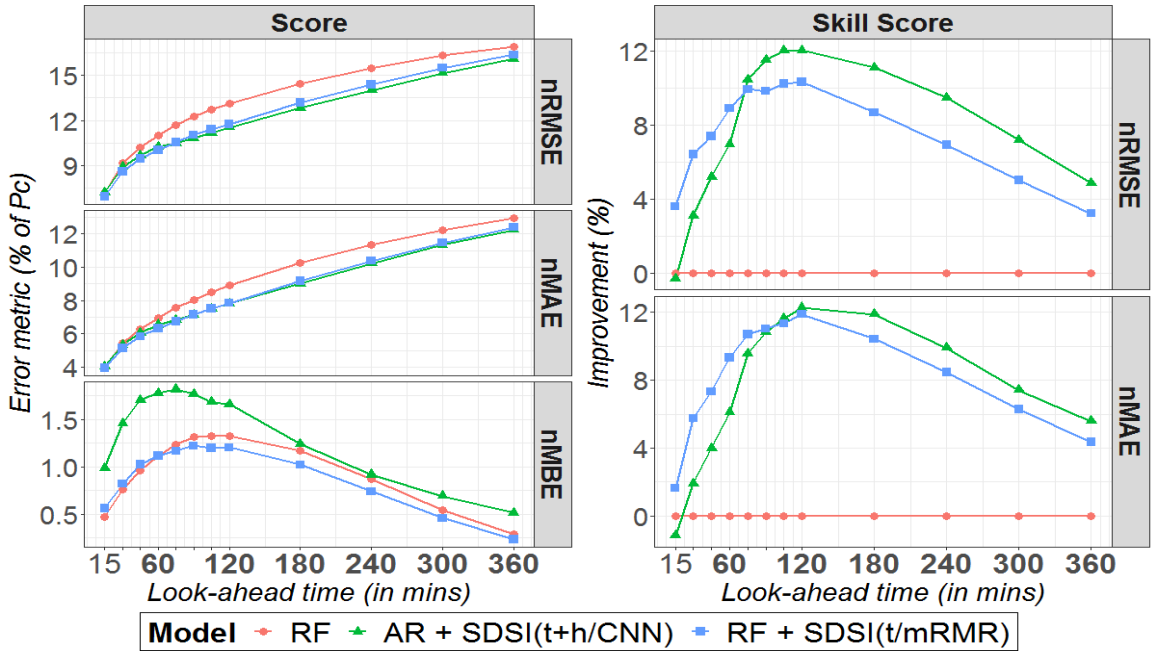



Figure 5.20 – Comparison of forecast accuracy obtained with best dimensionality reduction methods applied with AR and RF models.

To achieve higher accuracy for horizons greater than 1 to 2 hours ahead, the information contained in the last SDSI map observations is potentially not sufficient. We may need to look for additional features such as NWPs (e.g. wind direction and amplitude), which constitutes a potential way of improvement.

### 5.3.4 Comparison with a distributed network of PV plants

Now it is the time to compare the predictive power carried, in a way, by the SDU- and SDSI-based observations. Figure 5.21 and Figure 5.22 represent respectively the forecast accuracy of the AR and RF models fed with two sources of ST information.

It is obvious that satellite-based forecasts are better than SDU-based forecasts both in terms of nRMSE, and nMAE. Therefore, satellite-based observations offer an interesting option when the plant network suffers from a low density of units or a spatial distribution that does not match the wind distribution of the area. However, for the first three look-ahead hours, the SDU-based features allow the RF to achieve a slightly lower bias compared to the RF fed with SDSI. For that matter, in general, modelling strategies involving the RF models have a lower bias than those considering the AR models.

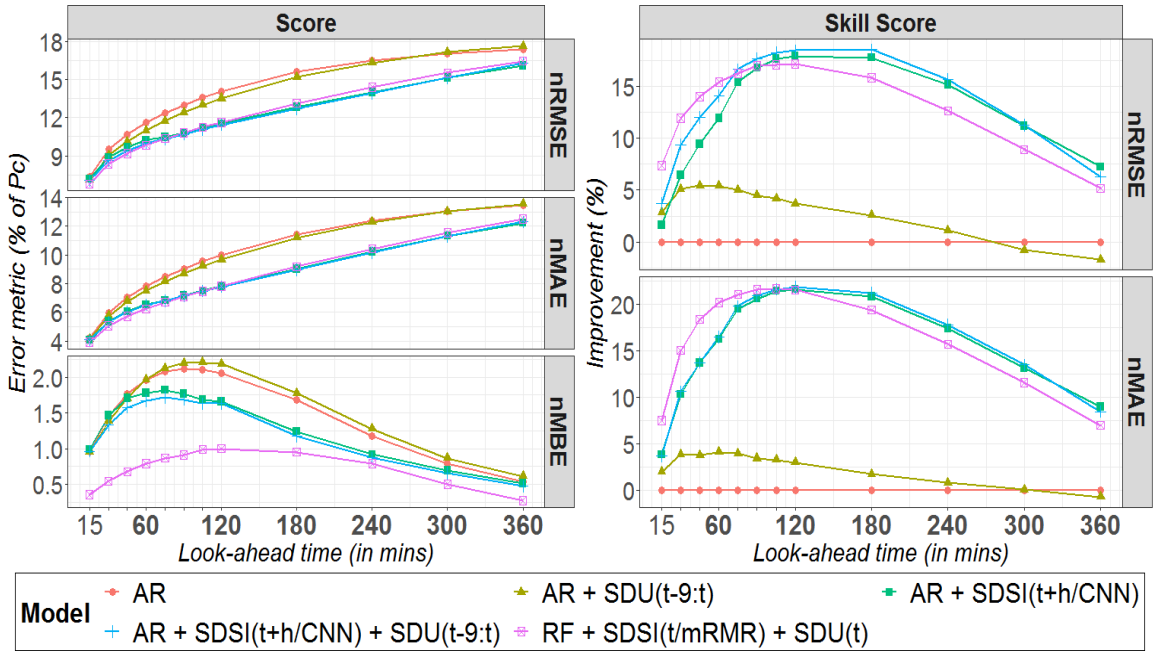



Figure 5.21 – Forecasting accuracy of the AR model fed with different sources of ST information. The ST version of the RF model is displayed for comparison purposes.

**Research Answer - *Mixing of data sources***

The simultaneous use of satellite and spatially distributed ground observations exhibits significant nRMSE improvements up to 1 hour and 2 hours ahead horizons with the AR and RF frameworks respectively. Beyond that, the information is mainly drawn from satellite measurements. This phenomenon may be explained by the fact that satellite-based observations offer a 360-degree view of the weather conditions in the vicinity of the plant, while SDU-based data are spatially dispersed. In addition, both sources of information possess different spatial resolutions; it is in the order of the $km^2$ for SDSI or the spatial size of the power plants in the case of SDU. In short, SDU offer very localised information w.r.t. SDSI.

## 5.4 Opacity maps

### 5.4.1 An under-represented source of information

The main limitation associated with classic satellite-based methods, such as the one used to derive SDSI observations, is that they only use visible spectrum channels. This provides daytime information but prevents the derivation of early morning forecasts due to a lack of time history. In this case, when forecasting models based on irradiance and/or PV production observations are launched before sunrise, no recent data are available, which

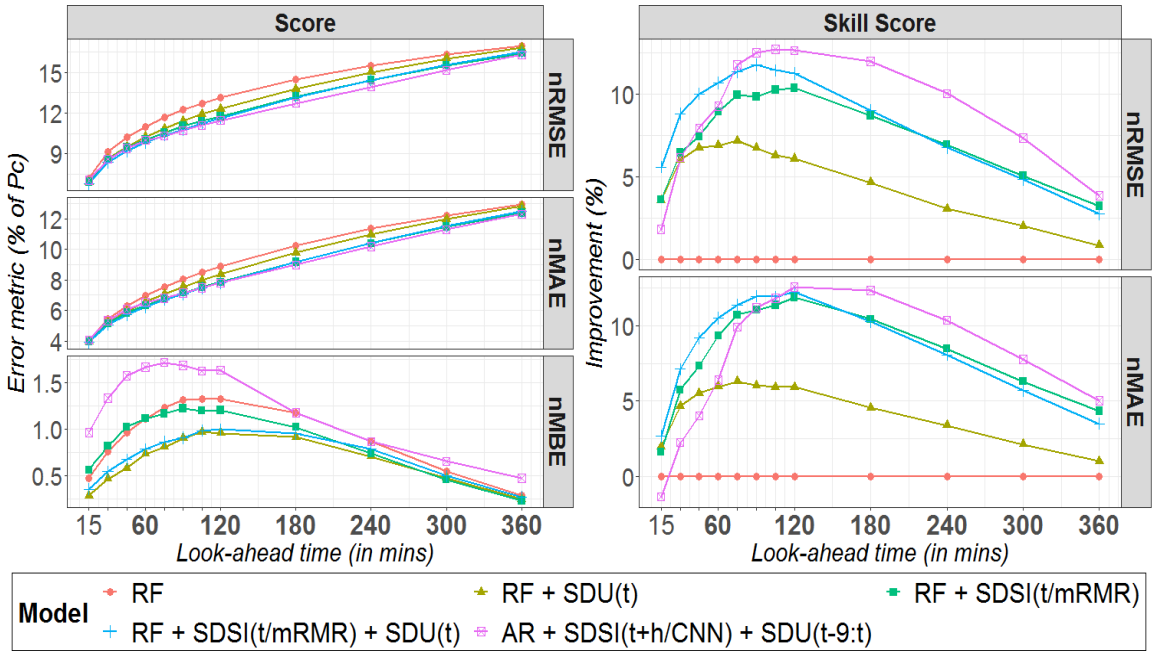



Figure 5.22 – Forecasting accuracy of the RF model fed with different sources of ST information. The ST version of the AR model is displayed for comparison purposes.

obliges the model to propose generic values learnt during its training. This explains the clear-sky profile observed at the beginning of the day in the forecast example displayed in Figure 5.23.

To improve forecasting accuracy for the early morning at least two strategies are conceivable: (1) considering NWPs, or (2) using nighttime observations of the cloud distribution. The latter option is made possible thanks to the development of observational satellites with infrared channels (which work day and night). In this regard, [114] provides a comprehensive description of how this technology can be used to derive the nighttime cloud index from the Spinning Enhanced Visible and InfraRed Imager (SEVIRI). In this section, we consider opacity maps (defined in Section 2.4.2.2). Such data allow the computation of forecasts at night so that they are available at sunrise (Figure 5.23).

In a context of operational forecasting, SDSI and opacity products are both generated from space observations but differ in the way the information is processed and expressed. Contrary to SDSI data, which are continuous variables, each pixel of opacity maps represents a categorical feature ranging from 0 (cloudless pixel) to 2 (pixel associated with opaque cloud), which intuitively drastically reduces the amount of information carried by the features.

**Research Gap - *Opacity maps compared with SDSI maps***

Thus, it is legitimate to wonder to what extent the information carried by both datasets is similar and whether one option prevails over the other and should be preferred to reduce financial costs.

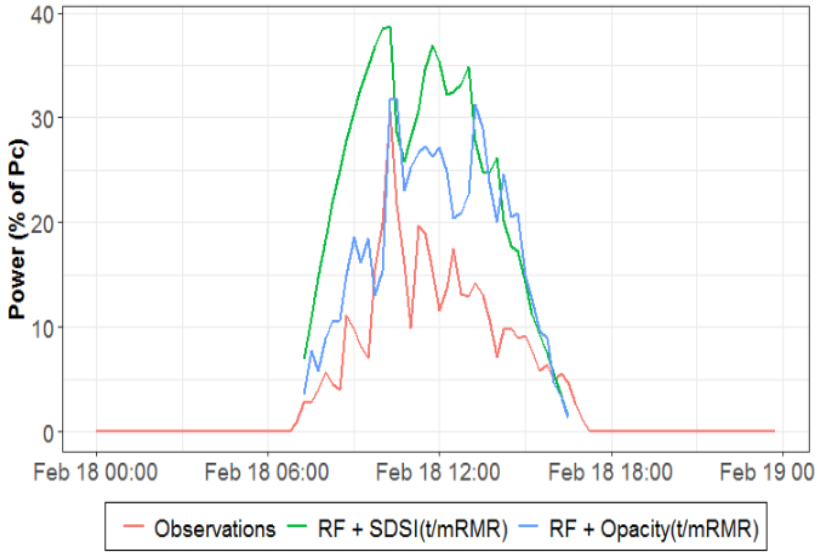



Figure 5.23 – Observations and 3-hour ahead forecasts of PV production at PV2. Forecasts are performed either with irradiance maps (i.e. daytime observations) or opacity maps (i.e. nighttime and daytime observations).

Surprisingly, to the best of the author's knowledge, only [114] and [104] consider infrared-based satellite maps to improve forecasting accuracy in the early morning (both studies propose CMV-based forecasting methods). Although PV production is very low during the very first moments of dawn, we observe a significant production rise in the first hours. This justifies developing research efforts to improve forecast accuracy, and all the more so as power systems can be pressured due to a morning ramp in electricity demand. Based on previous analysis performed on SDSI, we can stipulate that infrared satellite observations could be beneficial to forecast accuracy up to 6 hours ahead. A reasonable explanation to account for the under-representation of this type of input in the literature is that NWPs could represent a more interesting option. This leads us to the following research gap:

**Research Gap - *Opacity maps compared with NWPs***

What is the value of opacity map observations compared with NWPs of GHI for early morning forecasts?

### 5.4.2 Forecast accuracy

Based on previous results obtained with SDSI maps, we consider the RF model as the regression model and employ the mRMR feature selection process to extract relevant features from opacity maps. This approach is chosen for its easy implementation. The cloud-to-irradiance modelling is implicitly performed by the regression model.

A quantification of the forecast accuracy for the complete dataset [4] considering SDSI and opacity maps is displayed in Figure 5.24. We observe that in general the use of opacity-

4. In the sense that forecasts run during nighttime and daytime are analysed together.

based information improves forecast accuracy w.r.t. SDSI-based forecasts for all the horizons greater than 2 hours ahead, but provides rather similar nRMSE and nMAE scores for lower horizons. In addition, the information conveyed by the two types of input is complementary; proof of this is the improved performances obtained when considering both inputs simultaneously. The lowest normalised Mean Bias Error (nMBE) scores are achieved with models fed with opacity features.

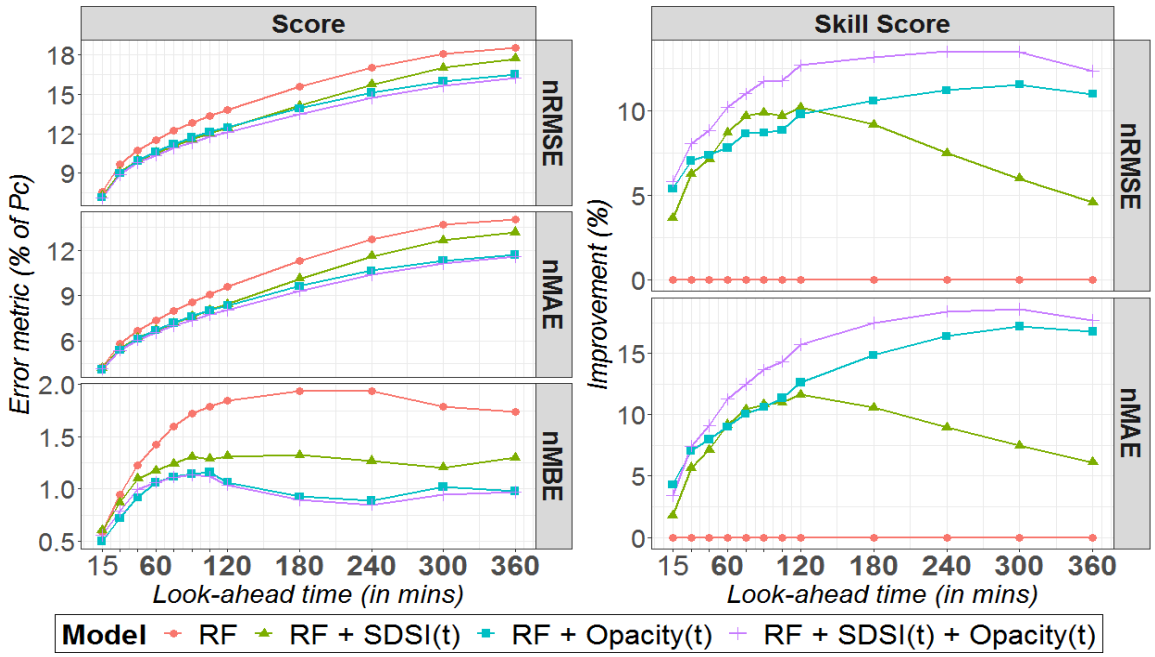



Figure 5.24 – Forecasting scores of the RF model fed with SDSI and/or opacity maps. Spatial information is preprocessed with the mRMR feature selection approach. Performances of models already analysed in previous figures (e.g. *RF + SDSI(t)*) may differ from performances observed here because of a change in the testing set due to the opacity maps' availability (Figure 2.9).

This graph demonstrates the added value of opacity inputs but it prevents distinguishing the respective impact of nighttime and ST information. To assess the influence of opacity maps over forecasting accuracy in detail, we compare two types of forecast: (1) forecasts generated during the night (i.e. relevant information is carried exclusively by opacity features and/or NWPs, while the latest irradiance-based observations are equal to zero), and (2) forecasts produced during the daytime (i.e. when the SDSI and/or the PV production observations have an impact on forecasting). On the one hand, Figure 5.25 shows that nighttime-generated forecasts clearly benefit from opacity features: improvement can reach more than 20% and 30% in terms of nRMSE and nMAE respectively compared to forecasts performed with irradiance-based information. We draw the reader's attention to the fact that the scores of the *RF + Opacity(t)* and *RF + SDSI(t) + Opacity(t)* models slightly differ because the addition of SDSI-based features alters the partitioning process performed by the RF model. On the other hand, when considering daytime forecasts, we observe that the impact of ST information conveyed by infrared observations on forecasting scores is am-

biguous. For very short-term horizons, the latter outperforms SDSI-based forecasts, while beyond the 1-hour ahead horizon, we observe a decrease in the forecast skill (the right-hand graph in Figure 5.25). This phenomenon may be explained, to some extent, by the limited number of classes used to encode opacity map pixels. However, as observed previously in Figure 5.24, both inputs carry diversified ST information, which is highlighted by the performance improvement resulting from the combination of inputs.

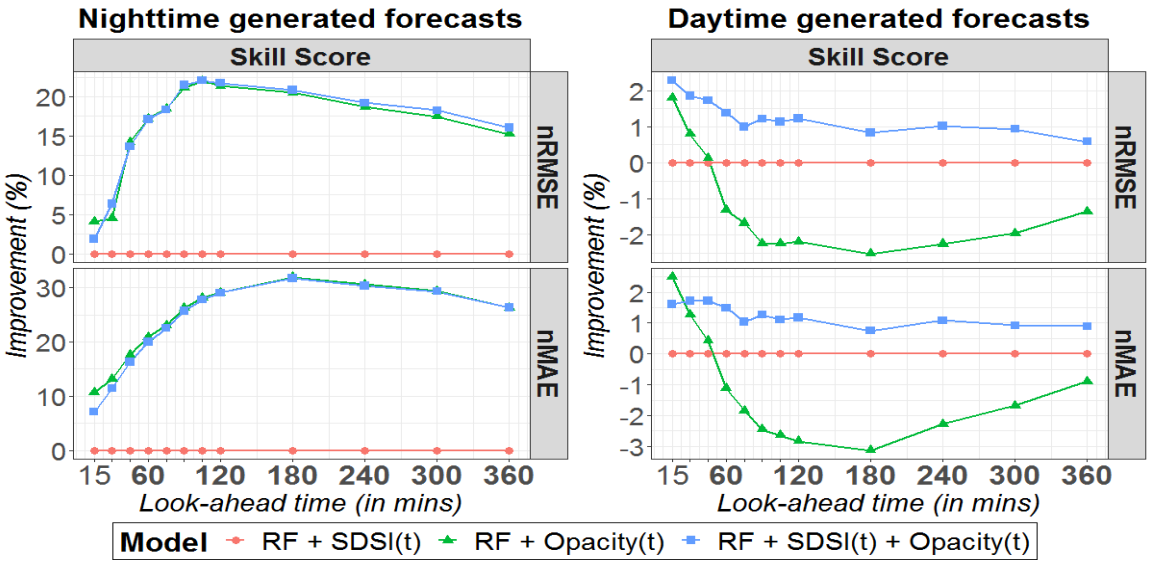



Figure 5.25 – Forecasting skill scores achieved with forecasts generated during nighttime and daytime w.r.t. the RF model fed with SDSI features.

**Research Answer - *Opacity maps compared with SDSI maps***

For daytime-issued forecasts, infrared-derived information proves to be relevant for very short-term horizons (i.e. from 15-min to 45-min). Beyond that, models fitted with this kind of input are outperformed by the same models fed with SDSI features by around 2% and 3% in terms of nRMSE and nMAE respectively. The highest forecasting performances are achieved with models trained on both inputs. In this case, we observed an improvement of around 1% compared with forecasts derived solely from irradiance information.

Figure 5.26 represents the forecast accuracy obtained with the complete dataset considering the GHI from the NWPs model and/or opacity map-derived observations. It shows that for the first horizons (up to +105 min), the forecasting performances achieved with opacity features (the *RF + Opacity(t)* model) are higher than those obtained with NWPs (the *RF + GHI(t+h)* model) in terms of nRMSE and nMAE. Beyond these forecast horizons, the GHI-based approach clearly prevails. For the 6-hour ahead horizon, the nRMSE scores achieved by the RF model fit on NWPs are around twice as good than those observed with opacity information w.r.t. an RF model only fitted on past PV production.

These high scores for high horizons result from two factors illustrated in Figure 5.27: (1) better forecasts for the early morning, and (2) accurate GHI predictions. For this kind of horizon, such results are consistent with the nature of the inputs: as the forecast horizon gets higher, the information carried by NWPs prevails over observations. We also observe that irradiance predictions and opacity features are complementary, and their combination gets the best from both approaches. In that regard, forecasts achieved with this last approach demonstrate the lowest bias.

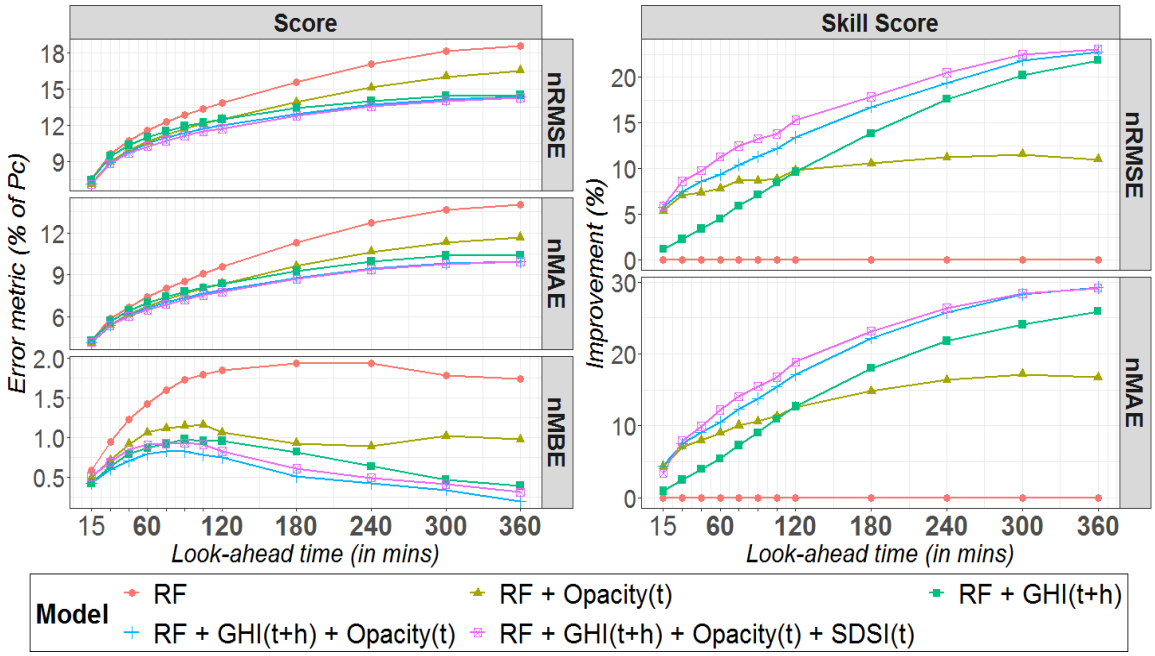



Figure 5.26 – Forecasting scores of the RF models fed with GHI forecasts and/or opacity-based features. Spatial information is preprocessed with the mRMR feature selection approach. We draw the reader's attention to the fact that performances observed in this graph are hardly comparable with figures from other sections because of a change in the testing set due to the opacity maps' availability (Figure 2.9).

If we focus on forecasts generated during the night for the early morning, we observe in Figure 5.27 that most accurate forecasts are provided by models based on NWPs independently of the considered horizon. Conclusions derived for daytime generated forecasts are similar to those obtained with the complete dataset.

**Research Answer - *Opacity maps compared with NWPs***

Thus, the added value provided previously by opacity information with respect to SDSI for nighttime-generated forecasts is eclipsed by numerical predictions of GHI (Figure 5.27). However, opacity maps still offer a clear advantage for very short-term forecasts generated during the daytime.

Therefore, considered separately, opacity data offer a clear advantage over SDSI observations for nighttime-generated forecasts. Yet, better performances are reached for the early

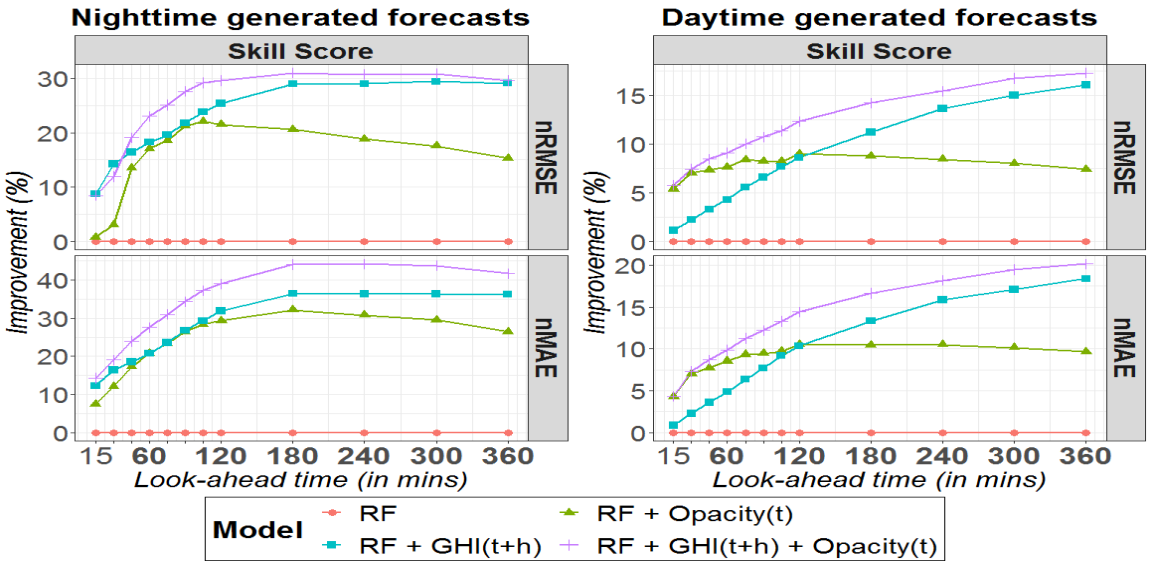



Figure 5.27 – Forecasting skill scores achieved with forecasts generated during nighttime and daytime w.r.t. RF models fed with past PV production observations.

morning when considering NWPs of GHI compared to opacity maps. In this respect, NWPs should be preferred. The ST nature of opacity maps improves daytime generated forecasts, but their informative potential is slightly lower than that of SDSI maps for horizons greater than 1-hour ahead. Thus, opacity maps appear as a variable able to compete with SDSI and NWPs features, but the aforementioned results highlight that it cannot replace them. Quite the contrary, we have shown that these sources of information are complementary and, when used together, they can improve forecasting accuracy. In that regard, Figure 5.26 demonstrates that the simultaneous integration of opacity and SDSI maps and GHI predictions provides the minimum nRMSE and nMAE scores, but that this gain is made at the cost of a slight degradation in the bias. Therefore, this proves the scientific interest of infrared-derived observations for the short-term prediction of PV power.

## 5.5 Conclusions

In this chapter, three sources of ST information have been investigated.

The integration of spatially distributed observations from nearby power plants exhibits accuracy improvements similar to what can be observed in the literature compared to a purely temporal approach. To reduce fitting time and overfitting issues, a time decorrelation distance is used to restrict the number of lags of ST features.

Limitations inherent to the spatial distribution and density of the power plant network under study lead us to consider SDSI. The most important impediment with this source of data is dealing with the problem of dimensionality and the associated intense computational costs. In this chapter, three dimensionality reduction methods have been investigated. The findings show that a nonlinear model such as RF is able to extract most of the relevant

information for the first hour-ahead horizons through a subset of a carefully selected set of features. To reach similar accuracy, a linear model needs to rely on a DNN structure to preprocess the information contained in the SDSI maps. Comparisons between SDU- and satellite-based forecasts highlight that the information contained in satellite-derived datasets is much richer despite a coarser spatial resolution. However, a combination of both sources of information can positively impact very short-term horizons.

Lastly, from a perspective of widening the range of available sources of spatial information, infrared-derived maps are studied. The findings show that their *spatio-temporal informative value* is somewhat limited compared to SDSI maps: for lead times higher than 45-min, satellite-based irradiance information prevails. In addition, despite a significant impact on the accuracy of nighttime-generated forecasts, the use of irradiance predictions achieves higher performances. However, a combination of the three datasets (i.e. SDSI, opacity maps, and NWPs) improves the RMSE and nMAE scores.

## 5.6 Résumé en Français

Avec le développement des sources d'énergies renouvelables, de nouvelles installations PV fleurissement un peu partout dans le monde. Cette expansion, couplée aux améliorations des moyens de télécommunication et des technologies de mesure, conduit à un nouveau paradigme centré sur les données spatio-temporelles. Les modèles qui en sont dérivés permettent de mettre à profit les dépendances pouvant exister entre des observations temporelles spatialement distribuées. Cette nouvelle classe de modèle de prévision permet, pour un producteur, de valoriser les mesures de production de son réseau de centrales sous forme d'une amélioration de la précision des prévisions. Ce paradigme ouvre également la voie à un futur marché d'échange de données.

### Observations de production spatialement distribuées

Dans ce chapitre, nous considérons un réseau de neuf centrales réparties majoritairement le long du Rhône. La distance inter-sites varie de quelques kilomètres à plus d'une centaine de kilomètres. Dans ces conditions, on imagine aisément que l'information relevée au niveau des sites les plus éloignés aura peu ou pas d'influence sur les performances prédictives. Ainsi, deux critères permettant de réduire le nombre de variables spatiales et temporelles sont considérés : (1) une distance seuil au-delà de laquelle les sites les plus éloignés ne sont pas considérés, et (2) un temps de propagation de l'information ST fini conduisant à un nombre de retards parcimonieux. Une étude préliminaire effectuée à la Section 4.4.2.5 du précédent chapitre met en avant l'inadéquation entre le régime de vent dominant auquel sont assujettis les différents sites et leur distribution spatiale. Pour y remédier, nous nous tournons vers une source de données plus flexible.

### Estimations de l'irradiance au sol par observation satellitaire

Contrairement aux réseaux de centrales de production, les observations d'origine satellitaire permettent de couvrir uniformément une zone géographique large avec une résolution spatiale de l'ordre de quelques kilomètres. L'augmentation du nombre de variables (e.g. une image avec un rayon de 100km est constituée de 900 variables) conduit à des problématiques d'augmentation du temps de calcul et de surapprentissage des modèles. Pour y remédier, trois approches de réduction de la dimensionalité sont analysées et comparées.

La première consiste à sélectionner un nombre prédéfini de variables. Typiquement dans la littérature nous trouvons des approches basées sur la maximisation d'un critère de corrélation (e.g. le coefficient de Pearson) par rapport à la variable cible, qui dans notre cas est la production PV à l'instant $t + h$. Néanmoins, nous mettons en évidence que ce type d'approche est loin d'être optimal dans la mesure où elle a tendance à sélectionner une information redondante caractérisée par une proximité spatiale des pixels sélectionnés. Pour remédier à cette lacune, nous considérons un autre schéma de sélection basé sur la minimisation de la redondance et la maximisation de la pertinence de l'information. A notre connaissance, cette approche n'a jamais été appliquée dans le domaine de la prévision de la production PV. L'utilisation de cette méthode permet la sélection de pixels spatialement distribués autour du site PV et conduit à une légère amélioration des performances prédictives.

Ensuite, nous avons considéré une analyse en composantes principales visant à réduire la dimension des données avant de nous tourner vers une approche basée sur un réseau neuronal convolutionnel dont l'objectif est de prédire l'irradiance au niveau du site PV à partir des dernières images satellite.

Une comparaison des différentes approches selon les deux modèles de régression considérés met en évidence que le modèle linéaire atteint de meilleures performances lorsqu'il est alimenté avec des prévisions de l'irradiance obtenues via le réseau convolutionnel. Le modèle non-linéaire, au contraire, est moins sensible au modèle de prétraitement utilisé. Enfin, nous démontrons que la précision des prévisions est meilleure en considérant des données satellite plutôt que des observations issues du réseau de centrale PV. Selon toute vraisemblance, ces conclusions sont spécifiques à notre cas d'étude et sont le fait de sa distribution monotone.

### Images d'opacité obtenues par canaux infrarouges

Enfin, dans une optique d'élargissement du spectre des données disponibles, la troisième et dernière partie de ce chapitre s'intéresse aux images d'opacité issues des canaux infrarouges des satellites. Ces données permettent d'accéder à la couverture nuageuse pendant la nuit et constituent donc une source d'information des plus pertinentes pour la génération de prévisions pour le petit matin. A notre connaissance, ce type de données est sous-représenté dans la littérature : uniquement deux références littéraires les utilisent ([104, 114]).

Dans un premier temps, nous confrontons les images d'opacité aux images d'irradiance.

Nous montrons que naturellement les images d'opacité ont un impact significatif pour les prévisions réalisées pendant la nuit (i.e. en l'absence d'observations récentes de l'irradiance ou de la production). Par contre, lorsque le modèle de prévision est lancé pendant la journée, le recours aux estimations de l'irradiance offre de meilleures performances pour des horizons au-delà d'une heure (nous observons des améliorations de l'ordre de 2%/3% en termes de nRMSE/nMAE).

Dans un second temps, les données d'opacités sont analysées vis-à-vis des prévisions numériques de l'irradiance. Dans ce cas, l'avantage des cartes d'opacité concernant les prévisions générées pendant la nuit est éclipsé par les prévisions de l'irradiance, celles-ci maintiennent néanmoins une influence notable pour les premiers horizons de prévision. Dans les deux cas, les meilleures performances sont atteintes en considérant simultanément les deux types de données, ce qui souligne la complémentarité des images d'opacité, d'irradiance et des prévisions numériques.

# Chapter 6

# Conditioned Learning

> In the spring I have counted one hundred and thirty-six different kinds of weather inside of four and twenty hours.
>
> Mark Twain (1876)

## Contents

## 6.1 Introduction

The previous chapter investigates the use of spatially distributed observations as a means to improve short-term forecasting accuracy. The model parameters are derived during the training step and used unchanged during the testing phase. Therefore, the model fitting mainly reflects the predominant situations encountered during its training stage. In an Spatio-temporal (ST) context, such an approach can be detrimental inasmuch as relevant spatially distributed features are not selected appropriately (e.g. in the case of a situation with adverse winds, only features in line with the prevailing winds are considered). To tackle this issue, we explore a dynamic modelling approach (i.e. which updates the model parameters), instead of models with fixed parameters. In this paradigm, rather than training models on the complete dataset, models are trained on a batch of data sharing common characteristics with the situation to predict. This leads to adaptive models. Special attention should be paid to have sufficient data to fit the models. So far, we have explored several options to improve or to diversify the information supplied to the regression models (e.g. clear-sky normalisation, use of satellite-derived observations). In this chapter, the main objective is to derive the mathematical foundations of a generic methodology to dynamically update model parameters.

The general workflow of this chapter is displayed in Figure 6.1.

## 6.2 Integration of information within forecasting models

### 6.2.1 Integration of weather information

Photovoltaic (PV) generation depends on a number of meteorological variables such as irradiance, cloud cover, airflow motion, ambient temperature and even air humidity. The combinations and interactions of these variables lead to a large range of weather states associated with significant varied dynamics. For this reason, Numerical Weather Predictions (NWPs) provide valuable information to Photovoltaic Production Forecasting (PVPF) models on the expected atmospheric state and how it will influence PV production. The predicted weather information can be integrated in the PVPF modelling chain in two different ways: either explicitly (i.e. with additional explanatory features) and/or implicitly (i.e. as state variables), which makes it possible local modelling.

#### 6.2.1.1 Explicit integration

The most straightforward method considers NWPs as additional explanatory features in the PVPF model (i.e. data are added linearly to the model). Only one model is fitted for a large range of weather situations thanks to the atmosphere dynamics explicitly carried by NWPs. This is a computationally inexpensive and easy way to include this type of information in PVPF models. Several references in the literature (e.g. [55, 61]) highlight

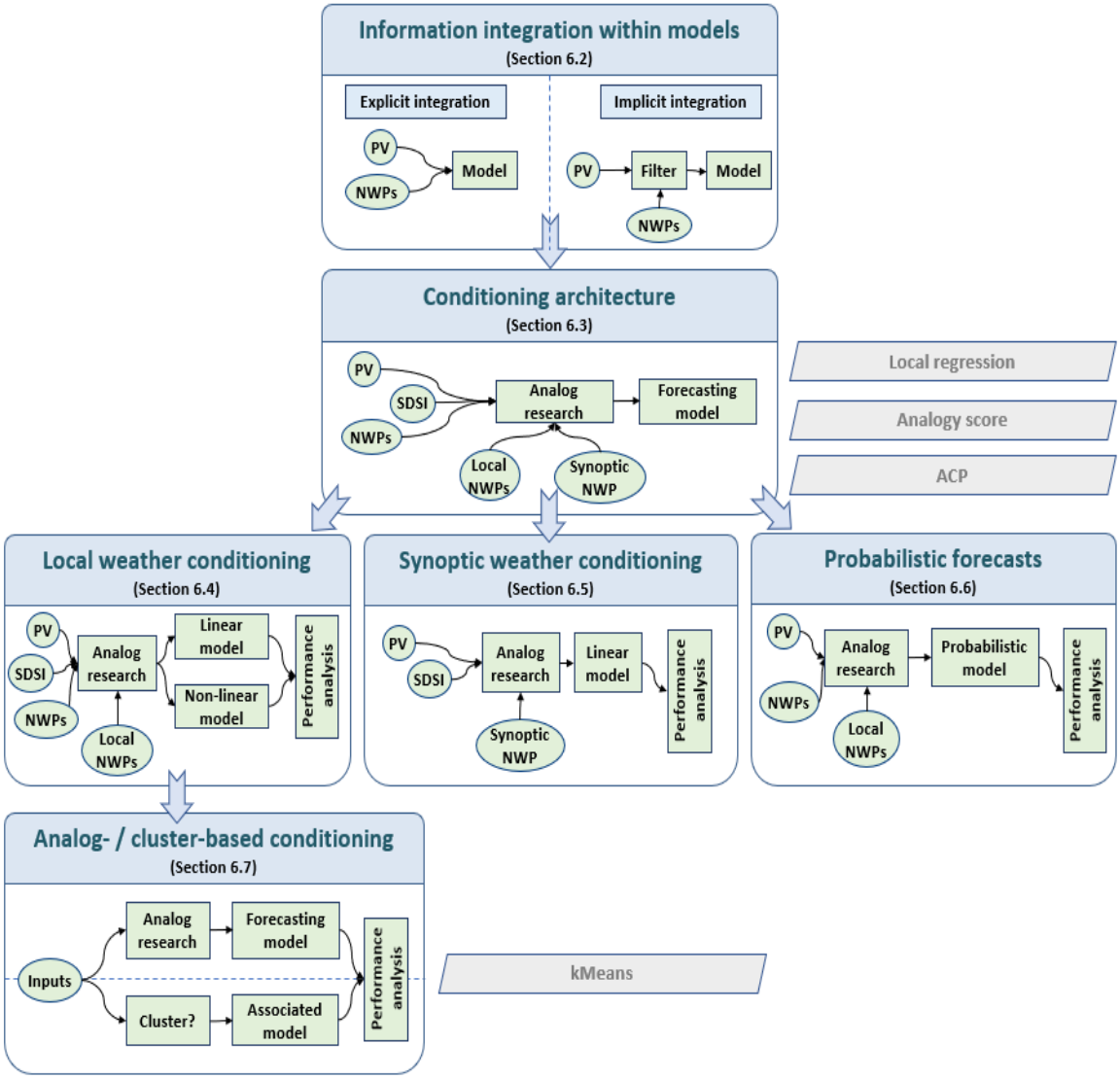



Figure 6.1 – General workflow of the chapter.

that the use of NWPs as regressor features improves short-term forecasting performances in comparison with models fitted only with past production observations. The integration of NWPs as inputs is beneficial as they inform on the tendency of future weather conditions.

#### 6.2.1.2 Implicit integration

The alternative paradigm is to consider the weather information as a state variable. Then, it acts as a kind of classification tool that gathers PV production data observed under similar atmospheric states. The assumption behind this approach is that similar PV production dynamics are observed under similar weather dynamics. From a mathematical point of view, weather information is included in a nonlinear way to perform local regression (i.e. the model is trained on a data subset that shares similar weather characteristics with the expected weather situation). To this end, a similarity metric must be defined to measure the likeness between two meteorological states. This approach provides a set of expert models

dedicated to specific atmospheric states and is adaptive in the sense that the training of the model is conditioned to the weather situation. Therefore, the atmosphere dynamics are implicitly carried by the PV production observations. It can be implemented in two ways: 1) either through a regime-switching model approach [64, 65, 255, 263], where each model is dedicated to a specific weather type (e.g. sunny or cloudy situations) or even through binning of weather variables [62], or 2) by taking a dynamic approach, where the model parameters are updated regularly [61, 66, 72].

The first option defines several pools containing situations with similar weather patterns; therefore, it can be viewed as a clustering-based algorithm. This classification can be based on weather types (e.g. sunny, cloudy, foggy, rainy) [64, 65] or on the binning of weather variables (e.g. [62] divides the learning set into eight groups depending on the wind direction). For each cluster, a dedicated model is trained. To provide a PV production forecast, it is necessary to determine the most representative cluster of the expected weather situation, and then to apply the corresponding model. In an operational context, this approach is relevant inasmuch as no model re-training is needed (a 98% reduction of the computational time is observed w.r.t. the dynamic approach described hereinbelow. More information is provided in Sections A and C.3), but forecasts may suffer from discontinuities (i.e. two successive forecasts may be produced with two distinct models).

The other option, which is investigated throughout this chapter, consists in training a new model for each situation based on the $N$ most similar past situations. In a space composed by the history of weather features, this approach searches the past situations that are the closest to the situation defined by the predicted features at time $t + h$. In that sense, this approach presents similarities to a k-Nearest Neighbours (kNN) algorithm. The main drawback is the need to re-train the model for each new forecast, which can be computationally expensive.

The literature proposes various terminologies to name these approaches: [62] proposes regime-based models, [65] describes its approach as a weather status pattern recognition model, while [264] bases its model on historical similarity. In an effort to unify these different approaches, we introduce the terminology *Weather-Conditioned (WHCO)* to refer to an approach, that operates a weather-based selection or classification in its learning dataset.

In the literature, the WHCO approach is applied to a large range of forecast horizons: from very short-term horizons [62], and short-term horizons [72], to day-ahead forecasting [63, 66, 263]. In [67] the authors go even further by developing a seamless model (i.e. a single model defined for several forecast horizons) which operates on a 5-min to 36-hour ahead horizons range, while [113] forecasts PV production up to 72-hour ahead.

This strategy offers the possibility of conditioning several types of forecasting models, such as Auto-Regressive with eXternal inputs (ARX) models [62], Artificial Neural Networks (ANN) [63], and Support Vector Machines (SVM) [64, 65]. To provide statistical informa-

tion regarding the uncertainty of the forecast, conditioning methods can be coupled with probabilistic approaches. In [66], the authors propose a Quantile Random Forest (QRF) model trained on the 30 most similar days, while [67] derives probability distributions by applying a weighted kernel density estimation model on the most similar PV production observations. It should be noted that [67] and [113] do not propose a WHCO approach as defined in the body of this paper: they use PV production observed under similar weather states to derive probabilistic laws, while here these data are fed into a regression model that infers the statistical relationship between the inputs and output.

From a performance perspective, it is worth mentioning that WHCO models exhibit greater forecasting skills than their counterparts trained on all past observations.

**Research Gap - *Comparison of integration modes***

The literature highlights that the WHCO strategy outperforms forecasting models trained on all past production observations. Similar conclusions are drawn when NWPs are considered as explanatory features. However, to the best of the authors' knowledge, no comparison has been performed to determine which approach provides the best forecasting performances. Therefore, which strategy leads to the most accurate forecasts?

**Research Gap - *Probabilistic forecasts***

The WHCO strategy allows us to train forecasting models on data subsets that share similar characteristics, namely production observed under similar weather states. In a way, this approach reduces the variability within the batch of datasets by focusing on observations that *matter*. Thus, in a probabilistic paradigm, can WHCO improve forecast attributes such as its reliability [a], and if so, its sharpness [b]?

*a.* Reliability assesses the statistical consistency between each class of forecasts and the corresponding distribution of observations.

*b.* Sharpness evaluates the concentration of the predictive distributions.

### 6.2.2 Integration of spatio-temporal information

In a similar way, ST information can be integrated in a PVPF model as either explanatory or state features.

On the one hand, the first option is the most common approach and can be based on several sources of ST information: [57, 58] use the PV production measurements of spatially distributed units, [55, 72] consider a selection of pixels derived from satellite imagery, while [56] fits solar forecasting models on observations from nearby irradiance sensors.

On the other hand, to the best of the authors' knowledge, only [67] considers ST information (namely Satellite Derived Surface Irradiance (SDSI) features) as state variables. This study considers variables from NWPs, in situ measurements, clear-sky profiles, and SDSI observations to identify situations with similar PV production dynamics. It is worth mentioning that in the scope of the paper, ST information is a means to improve the degree of similarity between analog situations.

In the context of WHCO, spot NWPs data are usually used (e.g. weather parameters predicted at the nearest grid point of the plant's location). Such data allow us to work with few inputs that mainly reflect the temporal evolution of local weather conditions without providing information regarding their spatial characteristics (e.g. cloud distribution). As a result, it seems difficult for the WHCO strategy to efficiently take advantage of ST information. To fill this gap, [62] considers a WHCO approach based on wind direction forecasts to select relevant ST information used as explanatory features. The authors propose a solar forecasting model for 10-second ahead conditioned by the wind features and fed with spatially distributed irradiance observations and past measurements at the location of interest. This study highlights that ST forecasting models benefit from a WHCO approach based on features related to cloud motions. Indeed, wind-conditioned forecasting models are able to select geographically distributed sensors in line with the cloud displacement, while un-conditioned models only select sensors in the direction of the most dominant winds.

**Research Gap - *Gridded data***

The approach proposed in [62] is valid if the cloud motion remains linear, which can be assumed for very short lead times but can be contested for higher periods of time and high spatial scales. In the context of precipitation forecasting, large-scale circulation patterns represented by geopotential fields, namely gridded NWPs (i.e. two-dimensional data), are used as state variables [68] owing to their proven influence over cloud generation. Inasmuch as we can derive pressure gradients that drive air flow from high- to low-pressure regions from these fields, we can wonder whether they could be used to provide a set of observations sharing temporal and spatial consistency. In other words, can geopotential fields be used to derive sets of PV production measurements observed under weather situations that evolve likely both in time and space in order to improve forecasting performances of the model based on SDSI information?

To the best of the authors' knowledge, geopotential fields are mainly used in the meteorological forecasting field, with very few articles at the junction with the Renewable Energy Sources (RES) forecasting field. For instance, in [265], the authors developed a post-processing method that produces a regime-based confidence interval of point Global

Horizontal Irradiance (GHI) forecasts. Synoptic scale geostrophic wind [1] forecasts are used to characterise the likeness of a given local weather states (e.g. in the considered case study, the clear-sky GHI forecasts associated with easterly geostrophic flow are more likely to be accurate as cloudy conditions are unlikely). More recently, in connection with the wind generation domain, [266] proposes a probabilistic forecasting model of the daily surface wind speed distribution from the geopotential field at 500 hPa at timescales ranging from 15-days to 3-months. It is worth mentioning that in addition to providing relevant information at the local scale, NWPs of features related to large-scale circulation of the atmosphere are more accurate than NWPs of local features such as GHI or wind.

## 6.3 Proposed architecture for model conditioning

To fulfil the above-mentioned objectives, we need to characterise the interactions that exist between (i) the different ways of integrating weather information, (ii) the state features dimensionality (i.e. spot- or gridded-features), (iii) the nature of explanatory features (i.e. temporal or ST), and (iv) the model family considered (i.e. linear or nonlinear regression models).

To do so, a modular architecture allowing the inhibition or the activation of some specific mechanisms occurring in the forecasting chain is proposed. It is composed of four main building elements (Figure 6.2): (1) the WHCO block, (2) the forecasting block, (3) the state variables block, and (4) the explanatory features block. This architecture may be seen as a generic data-driven forecasting model enhanced by a physics-based conditioning approach, which enables the model to perform local regression with respect to the atmospheric state.

### 6.3.1 Weather state conditioning

#### 6.3.1.1 The meteorologist's perspective

In the meteorology field, the analogy principle stipulates that similar weather states can be observed throughout time. Perfect similarity is hardly attainable due to the atmospheric variability, but similar situations can be found when considering large datasets and limited geographical areas. This has led to the development of forecasting approaches derived from Analog-based Method (AbM).

This set of methods can take many forms. For instance, such methods can be used as a downscaling approach by assuming that similar large-scale phenomena induce similar local-scale phenomena. In the precipitation forecasting field, AbM are used to derive probabilistic relations between large-scale variables (e.g. geopotential fields), named predictors, and local-scale features (e.g. precipitation) denoted as predictands [267].

1. It is the theoretical atmospheric flow for which the Coriolis and pressure-gradient forces are in equilibrium. This flow can be derived from geopotential height fields (Equation 8-9 in [265]).

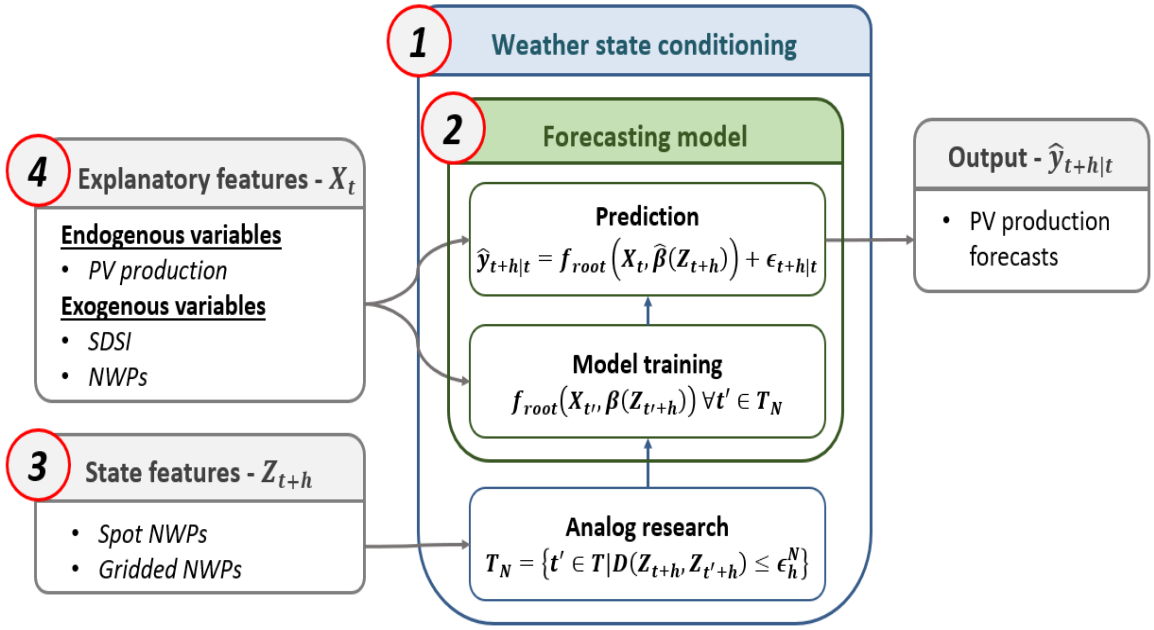



Figure 6.2 – Modular structure used to investigate the interaction between input integration strategies, input types, and model families.

In the present study, we assume that similar atmospheric states (i.e. predictor) lead to similar PV production dynamics (i.e. predictand). Thanks to the analogy principle, we can select a subset of past weather states similar to the future atmospheric situation. Instead of using the associated PV production observations subset to derive an estimation of the future production (this is performed in [67, 113]), here a dedicated forecasting model establishes the statistical relationship that exists between this subset and the associated explanatory features. This leads to weather-based expert regression models. Figure 6.3 illustrates how the AbM is used throughout this study. The modelling steps shown in the figure are as follows:

0. First, we build three datasets:
   (a) *The candidate archive* that contains weather forecasts,
   (b) *The response archive* that gathers the PV production observations,
   (c) *The explanatory archive* that represents the explanatory features dataset.
1. A score of analogy, $D$ (defined by Equation 6.2), measures the similarity between the target meteorological situation at time $t+h$ with past forecasts at lead time +h from the candidate situations archive, and ranks them. The $N$ most similar meteorological situations form the analog situations subset.
2. The $N$ associated PV production observations at lead-time +h are selected as well as the corresponding observations from the explanatory archive.
3. The selected elements from the response and explanatory archives are used to train a forecasting model, while last observations of the explanatory features at time $t$ enable

the generation of PV production forecast at time $t + h$.

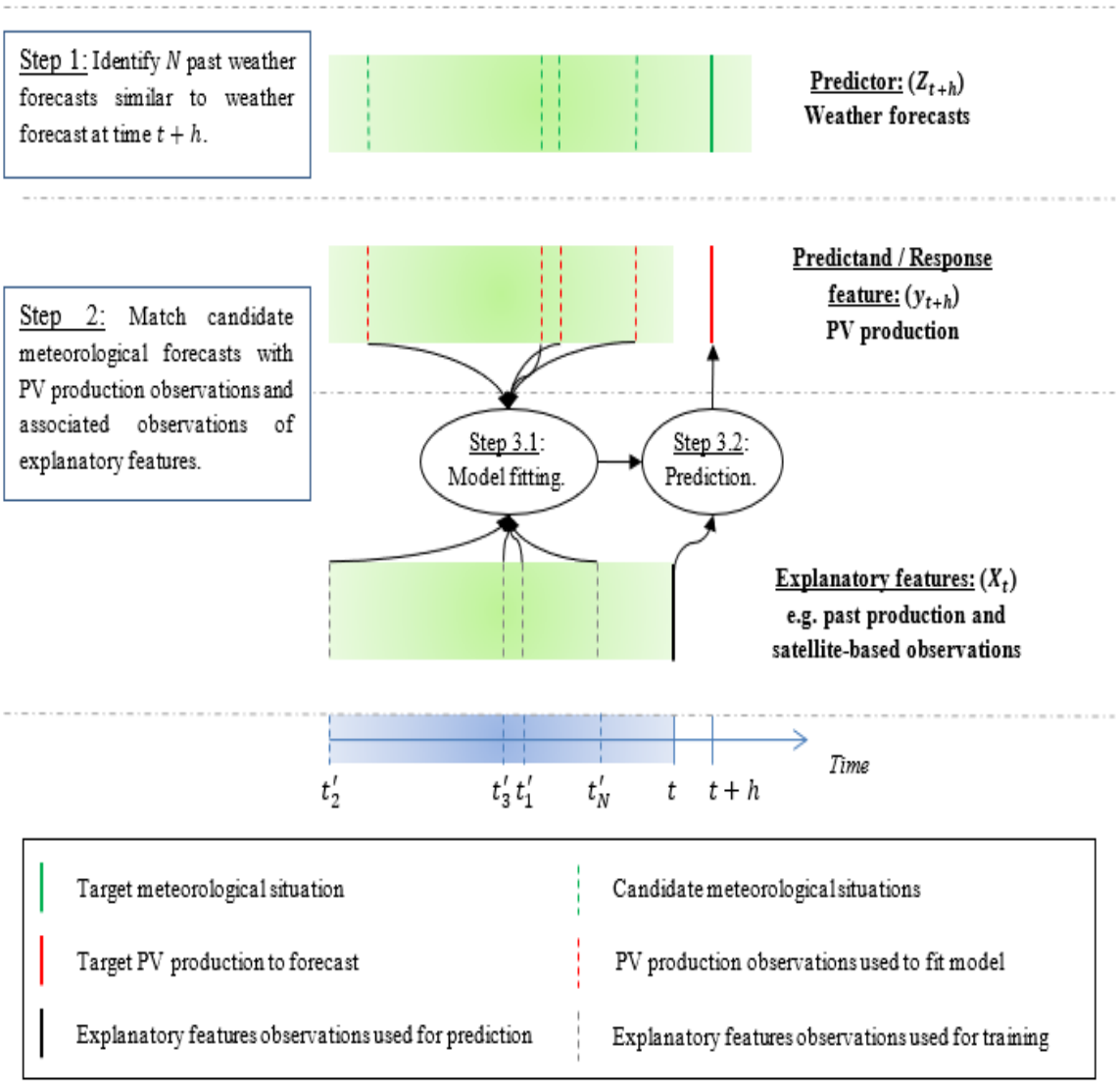



Figure 6.3 – Schematic representation of the training of the analog-based approach, inspired from [268, 269].

#### 6.3.1.2 Local regression

From a mathematical point of view, WHCO may be assimilated to a local regression approach [270]. Instead of fitting a regression model, denoted as *root model*, globally on the whole dataset of available observations $\mathcal{T}$, the fitting is performed locally on a subset $\mathcal{T}_N$ (Equation 6.1). This subset gathers $N$ observations associated with the neighbourhood of the focal point $Z_{t+h}$, namely the forecast of weather parameters. Attention is drawn to the fact that the fitting neighbourhood is defined within the state space (i.e. space containing state features, $Z_{t+h}$), while the model fitting is performed with explanatory features, $X_t$. This operation is repeated for all of the fitting points of the testing set in a rolling manner. We then obtain a dynamic architecture suited for online forecasting by updating the model

parameters regularly.

$$\hat{y}_{t+h|t} = f_{root}\left(X_t, \beta(Z_{t+h})\right) + \epsilon_{t+h|t}. \tag{6.1}$$

| | |
|---|---|
| $\hat{y}_{t+h\|t}$ | Response feature, |
| $f_{root}$ | Root regression model employed for the mapping of $X_t$ to $y_{t+h}$, |
| $X_t$ | Vector of explanatory features, |
| $Z_{t+h}$ | Vector of state features, |
| $\beta(Z_{t+h})$ | Vector of parameters to be estimated, |
| $\epsilon_{t+h\|t}$ | Error term from sources not considered. |

**Research Gap - *Weather conditioning and model family***

As shown in Section 6.2.1, the WHCO concept is already present in the literature and is applied to a wide range of forecasting models. Since the WHCO strategy can be viewed as a means to include nonlinear capabilities, what is its influence over linear and nonlinear models?

Equation (6.2) [113] presents the distance metric, $D$, used to measure the degree of similarity between the different observations of the state space, and to rank them according to their degree of likeness with the focal point. Lastly, only the $M$ closest elements are kept (Equation (6.3)). This score outperforms the traditional Euclidean distance thanks to the term under the square root, which takes into account the temporal evolution of the features. Thus, this approach retains weather situations that are locally alike (e.g. same irradiance level) and that evolve likely. In [113], the authors propose a grid search optimisation procedure to determine the optimal set of weights $\omega_i^A$, leading to the best forecasting performances. At this stage, it is important to point out that the use of the analog situations in our study differs from what is found in [67, 113]. In these studies, the Analog Ensemble (AnEn) method derives non-parametric probabilistic forecasts from the distribution of past PV production observations (e.g. in [67], the Probability Density Function (PDF) is generated from the analog situations through a Kernel Density Estimation (KDE)). Here, past observations of the explanatory and response features are used to train parametric models. In the present configuration, the grid search method is hardly conceivable due to the high computation costs induced by the Random Forest (RF) model fitting. Thus, we presume that the weights are uniform.

$$D(Z_{t+h}, Z_{t'+h}) = \sum_{i=1}^{N^A} \frac{\omega_i^A}{\sigma_i} \sqrt{\sum_{j=-\widetilde{t}}^{\widetilde{t}} (z_{i,t+h+j} - z_{i,t'+h+j})^2}. \tag{6.2}$$

$t$ Moment when the forecast is generated,
$h$ Lead-time of the forecast,
$t'$ Temporal observations from the learning set,
$i$ Index referring to the analog predictors,
$Z$ Vector of state features,
$z_{i,t}$ Element $i$ of the state vector $Z$ at time $t$,
$N^A$ Number of analog predictors,
$\omega_i^A$ Weight of analog predictors ($\sum_{i=1}^{N_A} \omega_i^A = 1$),
$\sigma_i$ Standard deviation of analog predictors,
$\widetilde{t}$ Half-width of the time window over which the metric is computed ($\widetilde{t} = 1h00$).

$$\mathcal{T}_N = \{t' \in \mathcal{T} \,|\, D(Z_{t+h}, Z_{t'+h}) \leq \epsilon_h^N\}. \tag{6.3}$$

$\epsilon_h^N$ Threshold distance used to retain a pre-defined number, $N$, of analog situations.

For information purpose, [66] use the Kolmogorov–Smirnov (KS) statistic as a similarity metric. This approach selects the 30 days that have the lowest KS distance between the Empirical Distribution Function (EDF) of the irradiance forecast for the day to be predicted and the EDF of the irradiance forecast for each day included in the database. This allows us to select days with similar weather dynamics and to shorten computing times (i.e. the same model is applied to all observations of the day).

#### 6.3.1.3 Number of analog observations

In the literature, it is common practice to use a fixed number of analog situations to reduce computational efforts. Yet, the number of observations needed to draw up a stable statistical law is expected to vary depending on the variability of the weather situation, or on the forecast lead time. In addition, a growing number of explanatory features and a few analog-based observations can lead to overfitting issues.

We introduce a selection procedure that select the optimal number of analog situations according to the forecast horizons, the site characteristics, the root model, and the number of explanatory features. This set of optimal analog situations is obtained through a grid search. The number of analogs associated with the forecasts that have the lowest normalised Root Mean Square Error (nRMSE) score is selected. This optimisation process is performed on the training dataset which is split into two subsets, one of which is dedicated to the model training (80% of the learning set) while the other is used for validation purposes. To account for seasonal effects, these two datasets are built in such a way that they contain the same proportion of data from the four seasons.

### 6.3.2 Forecasting models

The main element of this modelling chain is the root model which is employed to infer the statistical relationship between the response variable and the explanatory features.

Auto Regressive Integrated Moving Average (ARIMA) models [79] constitute a family of well-suited models to short-term PVPF [58, 61, 83]. Here, the ARX model (detailed in Section 2.2.3) is considered as the linear root model of our modelling strategy (Equation 2.2). The high number of available explanatory variables makes the model more complex and can undermine its accuracy. To tackle this issue, the Least Absolute Shrinkage and Selection Operator (LASSO) procedure [86] is implemented to perform feature selection and regularisation (Equation 2.3).

$$f_{root}\left(X_t, B^h\right) = \beta_0^h + \beta^h X_t^\intercal \tag{6.4}$$

$$(\hat{\beta_0^h}, \hat{\beta^h}) = \underset{\beta_0^h, \beta^h}{\arg\min} \left( \frac{1}{2} \sum_{t=1}^{N} \left( y_{t+h} - \beta_0^h - \sum_{j=1}^{P} \beta_j^h x_{t,j} \right)^2 + \lambda \sum_{j=1}^{P} \left| \beta_j^h \right| \right) \tag{6.5}$$

| | |
|---|---|
| $f_{root}$ | Root regression model employed for the mapping of $X_t$ to $y_{t+h\|t}$, |
| $X_t$ | Vector of explanatory features which may contain past production and satellite-derived observations as well as NWPs model outputs, |
| $B^h$ | Vector of the model parameters to be estimated, |
| $(\hat{\beta_0^h}, \hat{\beta^h})$ | Estimation of the regression coefficients, |
| $y_{t+h}$ | PV production at time $t+h$, |
| $\lambda$ | Hyper-parameter that determines the amount of shrinkage in the LASSO, |
| $(N, P)$ | Number of observations and variables. |

The second model considered is the RF model [91], which is a data-driven model able to perform nonlinear mapping between a set of input and output features. It is an ensemble learning method composed of several decision or regression trees grown in parallel, whose the outputs are averaged (Equation 2.4). Today, RF is one of the mainstream models employed in the field of RES forecasting: as an example, a recent forecasting competition was won by an architecture based on a QRF model [89]. More details are provided in Section 2.2.4.

$$\hat{y}_{t+h|t} = \frac{1}{T} \sum_{j=1}^{T} f_j(X_t) \tag{6.6}$$

| | |
|---|---|
| $\hat{y}_{t+h\|t}$ | Estimation of the response variable, |
| $f_j$ | j$^{\text{th}}$ regression tree. |

### 6.3.3 State variables

To select PV production data observed under similar weather patterns, it is necessary to work with weather parameters that accurately account for the PV generation process. Several approaches can be considered depending on the nature of the desired analogy. Spot data can identify analog situations on one particular location that evolves likely, but this data format does not guarantee that spatial patterns are preserved. We consider synoptic features (i.e. gridded-NWPs model outputs) to identify situations that evolve likely both in the temporal and spatial domains.

#### 6.3.3.1 Spot analogy

The features considered are the following outputs of the NWPs model: Surface Solar Radiation Downwards (SSRD), 2-m Temperature (T2M), and Total Cloud Cover (TCC) at the site position. These features are often used in the PVPF-related literature inasmuch as they directly affect PV production [113]. In addition, the solar azimuth and elevation angles ($\alpha^S$ and $\gamma^S$ respectively) (Figure 3.2) are added for two reasons: (1) despite irradiance-based explanatory features are normalised by clear-sky model outputs, [56] highlights that some periodical effects are still present in the normalised outputs; and (2) in a context of WHCO, these inputs enable us to implicitly take into account effects due to dawn (e.g. shading). In the spot analogy context, the vector of state features is built as:

$$Z_{t+h}^{\intercal} = \begin{bmatrix} SSRD_{t+h} \\ T2M_{t+h} \\ TCC_{t+h} \\ \alpha_{t+h}^{S} \\ \gamma_{t+h}^{S} \end{bmatrix}_{(5,N)} \tag{6.7}$$

#### 6.3.3.2 Synoptic analogy

**6.3.3.2.1 Geopotential fields** To account for the spatial and temporal evolution of the weather state in the PV farm's surroundings we consider the geopotential fields, which are synoptic scale features (i.e. scale of the order of $1,000$ km), as state features. Local weather parameters are directly influenced by synoptic pressure fields, which leads to the development of down-scaling techniques using AbM. These approaches assume that similar synoptic states lead to similar small-scales variables, which makes it possible to draw statistical relations between each scale. Geopotential fields are commonly used to forecast precipitation generation [68], and they demonstrate strong influence over wind direction. From the geopotential fields, we can derive the pressure gradient that drives the air flow from high- to low-pressure regions. Thus, geopotential fields are highly correlated with air flow and cloud generation, which makes them suitable to condition PVPF. Moreover, NWPs

models encounter some difficulty forecasting phenomena whose the governing processes occur at sub-grid scales, like explicit cloud formation, but turn out to be more reliable for forecasting large-scale atmospheric fields.

**6.3.3.2.2 Analogy score** In the literature specialised in precipitation forecasting, the score $S_1$ (Equation 6.8) [271] is usually employed to compute the degree of likeness between two situations when geopotential fields are considered. This metric is tailored for geopotential fields data: contrary to traditional analogy metrics which look for similarities point by point, this score compares the distance between the gradients of the target and candidate situations. This approach has been applied in the scope of PVPF in a previous study [72]. The main drawback of this score is to prevent the integration of additional state features.

$$S1 = 100 \frac{\sum_{i=1}^{I-1}\sum_{j=1}^{J}\left|\Delta_{i,j}^{i,T} - \Delta_{i,j}^{i,C}\right| + \sum_{i=1}^{I}\sum_{j=1}^{J-1}\left|\Delta_{i,j}^{j,T} - \Delta_{i,j}^{j,C}\right|}{\sum_{i=1}^{I-1}\sum_{j=1}^{J}\max\left(\left|\Delta_{i,j}^{i,T}\right|;\left|\Delta_{i,j}^{i,C}\right|\right) + \sum_{i=1}^{I}\sum_{j=1}^{J-1}\max\left(\left|\Delta_{i,j}^{j,T}\right|;\left|\Delta_{i,j}^{j,C}\right|\right)} \tag{6.8}$$

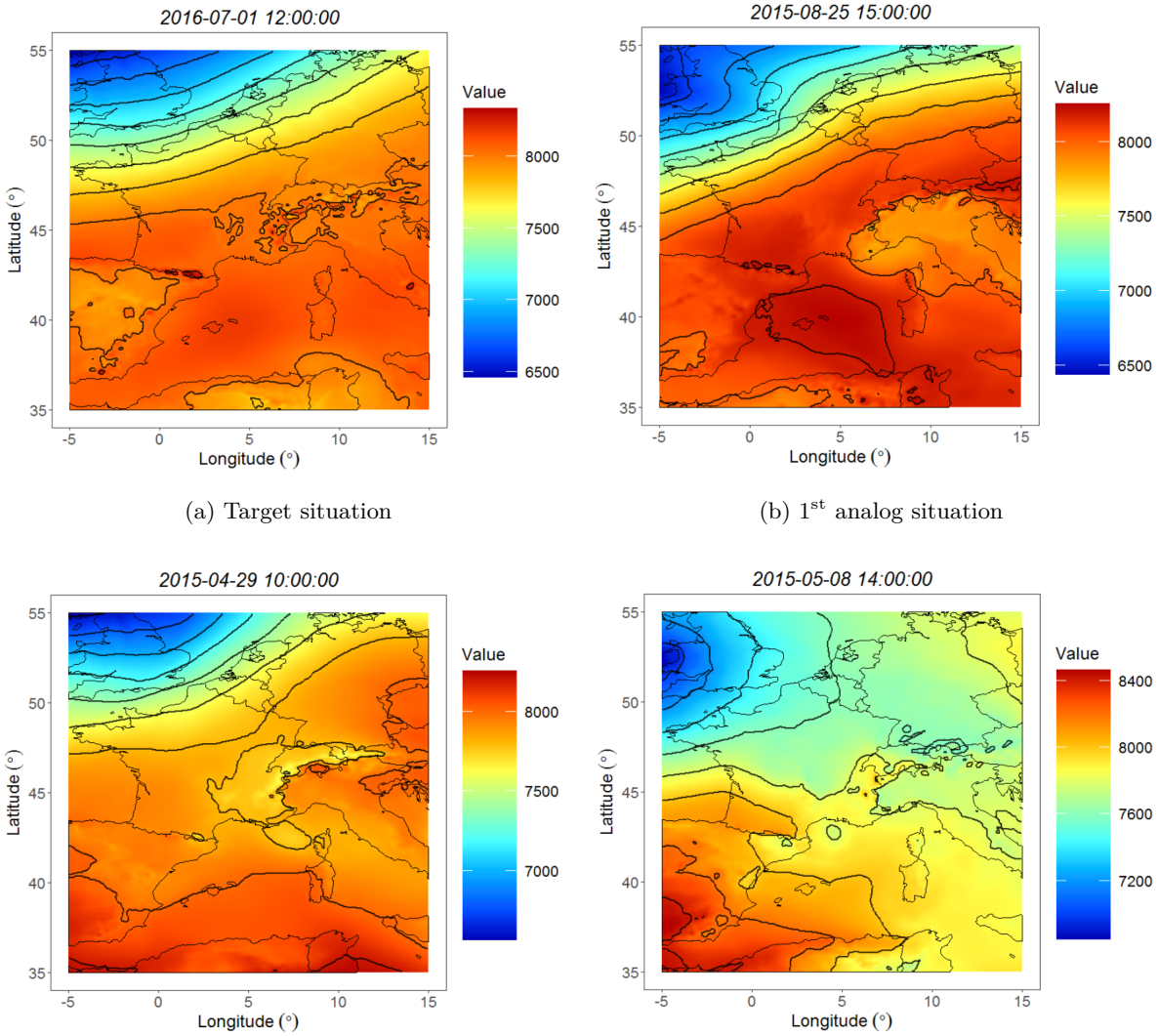
$$\text{Where: } \begin{cases} \Delta_{i,j}^{i,X} = V_{i+1,j}^{X} - V_{i,j}^{X} & X \in \{C,T\} \\ \Delta_{i,j}^{j,X} = V_{i,j+1}^{X} - V_{i,j}^{X} \end{cases}$$

| | |
|---|---|
| $T$ | Target situation (i.e. future state), |
| $C$ | Candidate situation (i.e. from past records), |
| $\Delta_{i,j}^{i,X}$ | East-west geopotential gradient, |
| $\Delta_{i,j}^{j,X}$ | North-south geopotential gradient, |
| $V_{i,j}$ | Geopotential field at grid node (i,j). |

The alternative explored in this study consists in performing a Principal Component Analysis (PCA) [252] to reduce the dimension of the state feature (methodology detailed in Section 5.3.2.2), and then to inject the projected data on the $N^{PCA*}$ Principal Components (PCs) into the analogy score $D$. Figure 6.4 shows a set of analog situations obtained with this approach when considering a geopotential field at 925 hPa. We observe that figures (a), (b), and (c) exhibit a north-south dipole configuration (i.e. a low-pressure area in the north and a high-pressure area straddling central Europe). The degree of likeness between the target situation and its $100^{\text{th}}$ analog is somewhat low. This is due to the high spatial extent of the analogy window and the low historical depth: only diurnal data from the training set are considered. This graph supports the proposed methodology to identify weather situations with similar spatial patterns.

A similar graph is proposed in Section C.1 that represents a set of analog situations obtained with the $S_1$ score and considering the same target situation. We observe that the $1^{\text{th}}$, and the $10^{\text{th}}$ analog situations are also very similar to the target situation. The main difference between Figure 6.4 and Figure C.1 is that the $100^{\text{th}}$ analog situation (Figure C.1d) obtained with the $S_1$ score visually possesses a higher degree of similarity with the target than the $100^{\text{th}}$ analog situation obtained with the coupling of the PCA approach and the

analogy score $D$ (Figure 6.4d). This may be explained by the fact that the last approach integrates solar angles as additional state features while the $S_1$ score is fed solely with the geopotential field.




(a) Target situation

(b) 1st analog situation

(c) 10th analog situation

(d) 100th analog situation

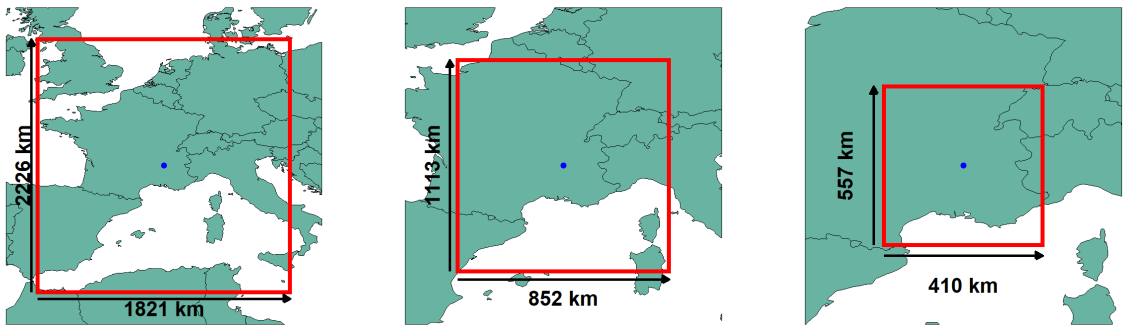
Figure 6.4 – Examples of analog situations (**b**), (**c**), (**d**) with regard to the target situation (**a**) obtained with the 925 hPa geopotential field by combining the PCA-based feature reduction approach and the analogy score $D$.

Thereafter, to enable a fair comparison with the spot conditioning, we also integrate the solar azimuth and elevation angles as a proxy of time. The resulting vector of state features is:

$$Z_{t+h}^{\intercal} = \begin{bmatrix} PC_{1,t+h} \\ \vdots \\ PC_{N^{PCA*},t+h} \\ \alpha_{t+h}^{S} \\ \gamma_{t+h}^{S} \end{bmatrix}_{(N^{PCA*}+2,N)} \tag{6.9}$$

**6.3.3.2.3 Choice of the pressure level and grid domain** PV generation is affected by various types of clouds evolving at different pressure levels. It is therefore important to pay special attention to the geopotential field considered. Geopotential fields at pressure levels of 500 hPa and 925 hPa are known to contain essential information about the dynamic and thermodynamic physical processes behind precipitation generation and distribution [272]. As geopotential fields are 2-dimensional features, it is also essential to demarcate the spatial range of valuable information.

Thus, a sensitivity analysis is carried out on the pressure levels of geopotential fields (500 hPa and 925 hPa) and three spatial windows centred on the Rhone Valley (Figure 6.5). We considered the Auto-Regressive (AR) model to assess the forecasting performances and the nRMSE and normalised Mean Absolute Error (nMAE) scores. This analysis highlighted that the best forecasting performances are reached with the 925 hPa pressure level and the spatial window (c) (Appendix C.2).




(a) Window 1. (b) Window 2. (c) Window 3.

Figure 6.5 – Grid domains used for analog research, (a): $\{N : 55°, W : -5°, S : 35°, E : 15°\}$, (b):$\{N : 50°, W : 0°, S : 40°, E : 10°\}$, (c):$\{N : 47.5°, W : 2.5°, S : 42.5°, E : 7,5°\}$.

### 6.3.4 Explanatory variables

The root models in Equation 6.1 are fed with two kinds of explanatory variables: either endogenous inputs (i.e. PV production) and/or exogenous inputs (i.e. spot NWPs and SDSI). Within the scope of short-term PVPF, endogenous inputs are essential, and consequently they are systematically integrated. In a next step, spot NWPs and SDSI features are considered individually or together to assess their influence on forecasting skills.

Equation (6.10) represents the regressor vector which contains all available inputs.

$$X_t^\intercal = \begin{bmatrix} P_{t-L:t} \\ SDSI_{t-L:t}^{1:N^{SDSI}} \\ SSRD_{t+h} \\ T2M_{t+h} \\ TCC_{t+h} \\ \alpha_{t+h}^S \\ \gamma_{t+h}^S \end{bmatrix}_{((L+1)\cdot(N^{SDSI}+1)+5,1)} \tag{6.10}$$

| | |
|---|---|
| $P_{t-L:t}$ | Last observations of PV production from lag $t-L$, |
| $N^{SDSI}$ | Number of satellite pixels determined with the minimal-Redundancy-Maximal-Relevance (mRMR) selection algorithm, |
| $SSRD$ | Surface solar radiation downwards at site position, |
| $T2M$ | 2-m temperature at site position, |
| $TCC$ | Total cloud cover at site position, |
| $\gamma^S$ | The Sun's elevation angle at site position, |
| $\alpha^S$ | Solar azimuth angle at site position. |

In Section 5.3.3, we perform a comparison between three information extraction methods from satellite-based maps. They reveal that in the case of the ARX model, the pre-processing of the ST information through a Convolutional Neural Networks (CNN) architecture leads to the best accuracy. Here however, the WHCO approach is coupled with the mRMR feature selection algorithm to investigate the potential benefits of their combined use. The mRMR algorithm is run on the whole training dataset to select a fixed subset of $N^{SDSI}$ features according to the forecast horizon $h$. As shown in Section 5.3.2.1.2, this procedure tends to select SDSI features in every direction. In Chapter 5, the models are trained on the whole training set which promotes predominant wind regimes in specific directions. In this chapter, quite the contrary, the root model parameters are updated regularly using the WHCO approach. This makes it possible to take into account the different wind regime distributions that the plant experiences over time, and it is expected to value the spatial distribution of SDSI features.

### 6.3.5 NWPs model outputs

#### 6.3.5.1 Temporal granularity

The explanatory variables and the PV power forecast outputs have a 15-min granularity, while the state variables consist in hourly predictions. Instead of performing expensive temporal interpolations of the state variables at a 15-min time-step, we assume that the atmospheric state remains constant from time $t-00h15$ to time $t+00h30$. Thus, when two situations, $t_1$ and $t_2$, are considered to be similar, the situations ranging from $t_1-00h15$

to $t_1 + 00h30$ and the situations ranging from $t_2 - 00h15$ to $t_2 + 00h30$ are also considered alike.

#### 6.3.5.2 Runs of NWPs model

NWPs models are computed several times a day. These sets of forecasts are named *runs*. Depending on the lead time, several predictions can be issued for the same time according to the NWPs run considered (e.g. predictions for time 13:00:00 UTC can be provided by the runs of 00:00:00 UTC and 12:00:00 UTC on the same day). As a result, two approaches are considered according to the weather information integration strategy.

First, we may consider that each run has distinctive features: the number and position of initial observations used to initialise the numerical model vary according to its launching time, which may impact the quality of the forecasts. Therefore, when NWPs are considered as state features, it is relevant to compare predictions with similar errors. To do so, we consider runs delivered at the same time of the day when looking for analog situations.

Alternatively, we may focus on the fact that forecasting precision tends to decrease as the lead time increases. Thus, when NWPs are considered as explanatory variables, predictions from the most recent run are considered.

### 6.3.6 Considered architectures and terminology

To assist the reader in understanding the configurations assessed, Figure 6.6 gathers model denominations as well as block diagrams representing the model architectures:

- $Model \in \{AR, RF\}$. The AR and RF models are investigated.
- $X_1, X_2 \in \{\emptyset, NWPs, SDSI\}$. Forecasting models can be fed with PV production observations and spot NWPs and/or neighbouring satellite pixels obtained by the mRMR-based method detailed in Section 5.3.2.1.2.
- $Z \in \{Spot, Gridded\}$. PV production forecasts are conditioned either with spot- or gridded-NWPs (they are denoted respectively as *local* and *synoptic* WHCO).

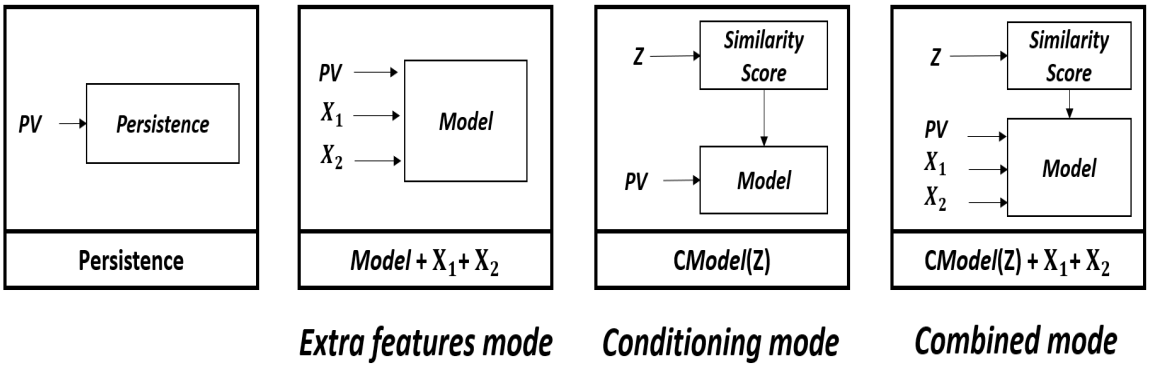



Figure 6.6 – Models designation and corresponding structures. *CAR* and *CRF* terminologies stand for Conditioned-AR and Conditioned-RF.

## 6.4 Local weather information

### 6.4.1 Optimal number of analog situations

In this paragraph, we evaluate the grid search optimisation framework introduced in Section 6.3.1.3. To do so, we compare the forecasting performances of the *CAR(local)+SDSI* model based either on a fixed number of similar weather situations or on the optimisation framework. We choose a forecasting model fed with ST inputs to generate a configuration with a high number of variables and to place ourselves in a situation prone to overfitting. Figure 6.7 highlights that the optimisation framework improves forecasting performances for very short-term horizons in comparison with the fixed analog number approach.

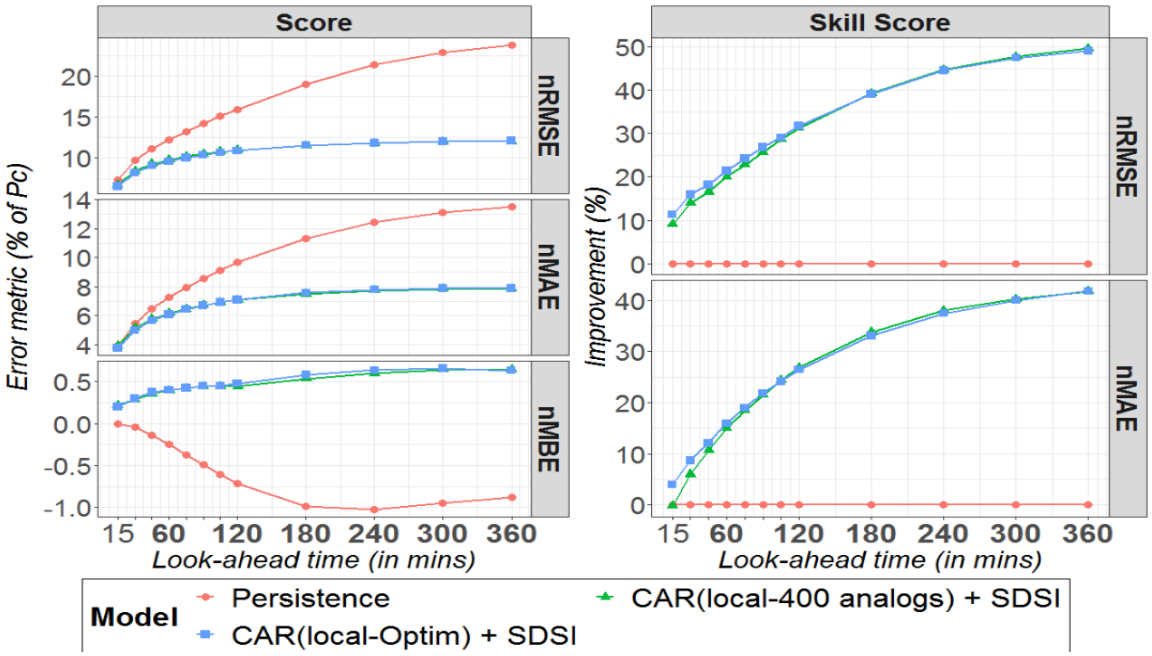



Figure 6.7 – Influence of the grid-search optimisation process on forecasting performances in comparison with a model trained on a fixed number of analog situations. The grid-search algorithm compares the performances of models trained with $N = \{200, 400, 800, \ldots, 4000\}$ analogs.

Figure 6.8 illustrates the variation of the optimal number of observations obtained with the grid search algorithm according to the look-ahead time. We observe that the ST approach (i.e. *CAR(local)+SDSI*) requires more observations during its training than its temporal counterpart fed only with past production observations (i.e. *CAR(local)*). Thus, a wider dataset (i.e. with more explanatory features) requires a deeper structure (i.e. with more observations) to infer relevant statistical laws. Moreover, the figure shows that the number of analog situations required during the learning phase decreases with the forecasting horizon. This phenomenon may be explained by the fact that as the lead time increases, the uncertainty regarding the future also increases which constrains the forecasting model to focus on the most similar situations to derive relevant statistical laws.

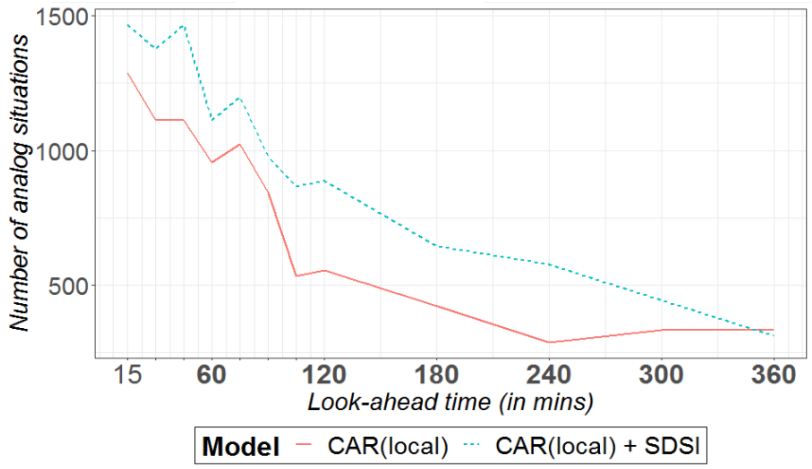



Figure 6.8 – Averaged number of analog situations determined by the optimisation framework for the 9 PV farms under study in a temporal or an ST context.

### 6.4.2 Weather information integration in a linear model

#### 6.4.2.1 Nonlinear dependent feature

The WHCO approach is a straightforward and efficient way to integrate explanatory features that have a nonlinear relationship with the response variable in a linear model. To illustrate this statement, we consider the integration of the azimuth angle, $\alpha_S$, in the ARX model. Figure 6.9 shows that performances achieved by considering the azimuth angle as an explanatory feature (i.e. the *AR + Azimuth* model) are outperformed by the state feature integration mode (i.e. the *CAR(Azimuth)* model).


Figure 6.9 – Integration of the solar azimuth angle either as an additional explanatory feature or as a state feature in an AR-based forecasting model.

#### 6.4.2.2 Two complementary approaches

Figure 6.10 represents the forecasting performances achieved by the ARX model fed with the SSRD feature. The WHCO approach (i.e. the *CAR(SSRD)* model) performs poorly compared to its counterpart (i.e. the *AR+SSRD* model). When the two integration modes are simultaneously employed (i.e. the *CAR(SSRD)+SSRD* model), resulting performances are the highest for both skill scores considered. Therefore, the extra-feature mode should be preferred for features that have a linear dependence on the response variable. Yet, far from being two opposed integration modes, the WHCO and the extra-features strategies assess different kinds of information which can complement each another.

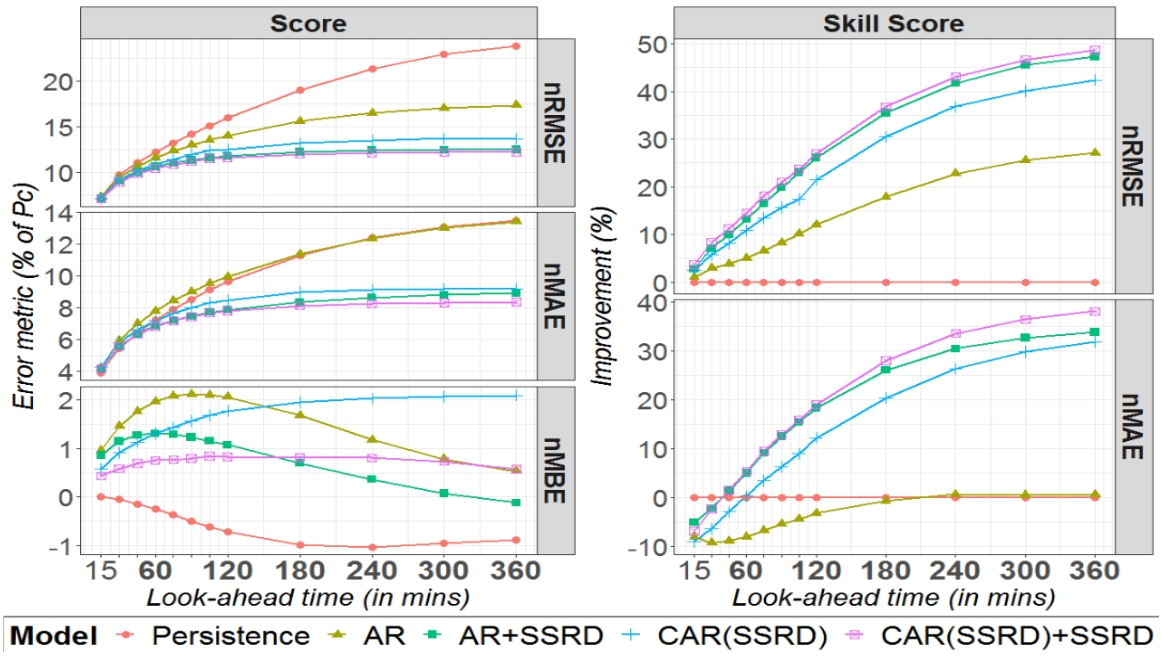



Figure 6.10 – Influence of the feature integration approach of the SSRD (i.e. as explanatory feature, state feature or both) on forecasting performances.

**Research Answer - *Comparison of integration modes***

In the case of a linear regression model, the WHCO approach makes the best of features that have nonlinear dependencies on the response variable in comparison with a straightforward integration as explanatory features. Nonetheless, features with a linear dependence provide better performances when considered as explanatory features, but higher scores can be reached when both modes are simultaneously employed. These findings are valid for a linear model.

### 6.4.3 Interaction between forecasting model families, sources of information, and integration strategy of weather data

In this section, we compare the different forecasting architectures to determine the best way of integrating data and obtaining optimal forecasting performances. Figure 6.11 and Figure 6.12 gather the forecasting skill scores of the $AR$ and $RF$ models coupled with the past PV production observations, and/or NWPs and/or SDSI observations. The conditioning of these models on local NWPs is also evaluated.

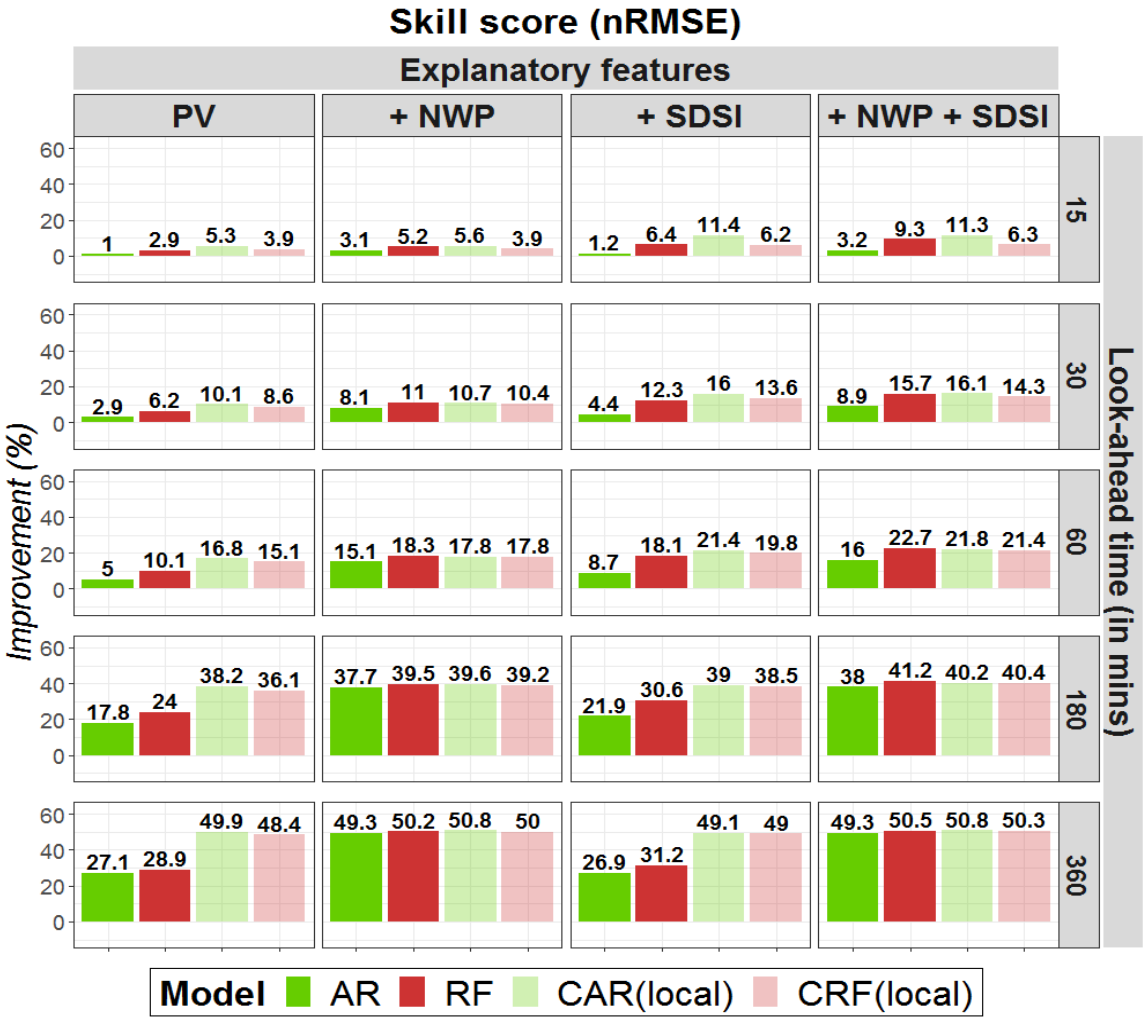



Figure 6.11 – nRMSE skill scores with regard to the persistence model. Dark colours symbolise forecasting models trained on the whole dataset, while light colours stand for WHCO models. Columns represent the explanatory features, while rows indicate the lead time of the forecasts. The number above the bars indicates the exact value of the improvement metric.

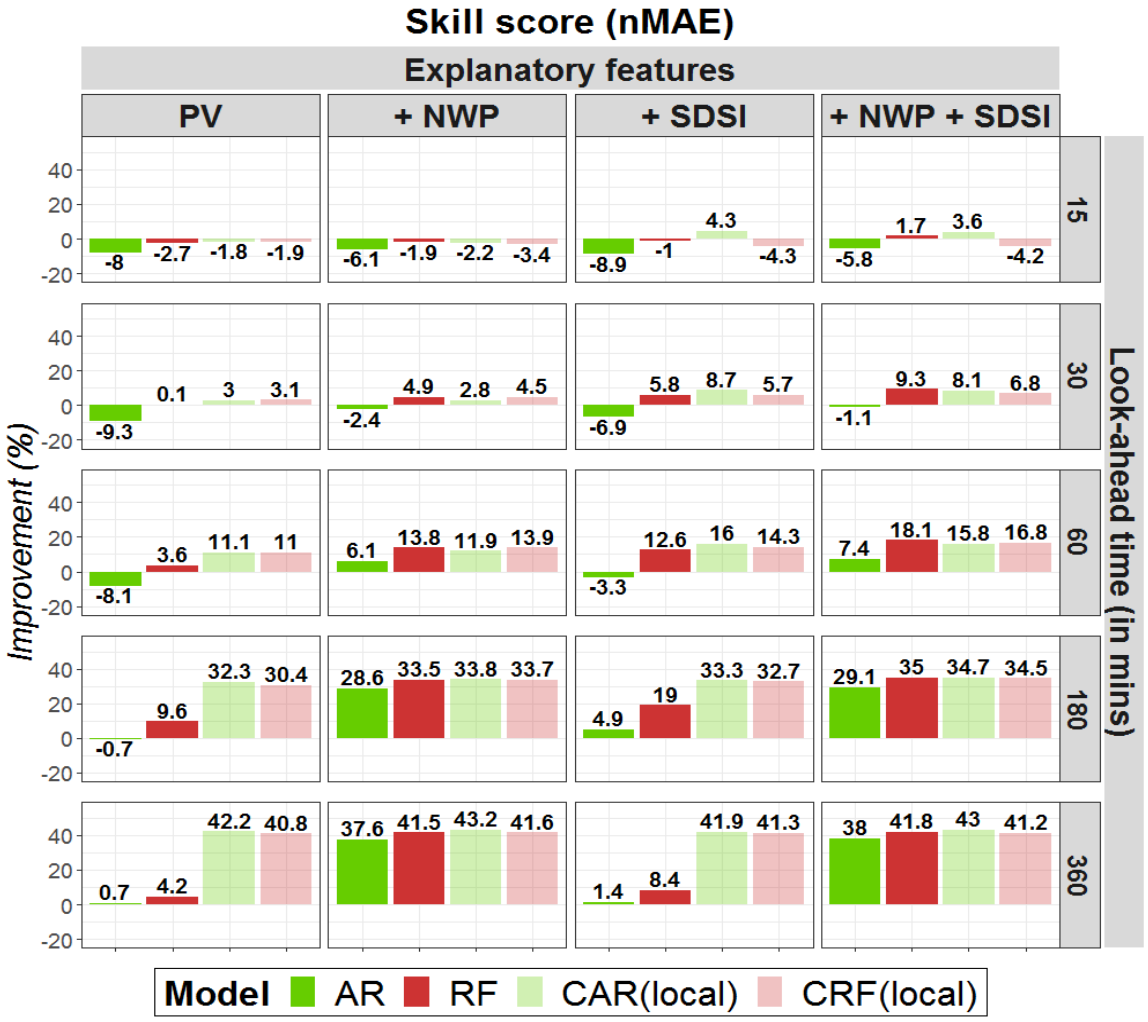



Figure 6.12 – nMAE skill scores with regard to the persistence model. Dark colours symbolise forecasting models trained on the whole dataset, while light colours stand for WHCO models. Columns represent the explanatory features, while rows indicate the lead time of the forecasts. The number above the bars indicates the exact value of the improvement metric.

#### 6.4.3.1 PV production observations

When only PV production observations are available [2], the nonlinear model turns out to be more accurate than the linear model, both in terms of nRMSE and nMAE. This observation is valid for all of the forecast horizons under study.

#### 6.4.3.2 PV production observations + SDSI

In an ST context , the *RF+SDSI* model outperforms the *AR+SDSI* model for all considered horizons and metrics. A comparison of the skill scores between temporal-based forecasts (i.e. the *AR* and *RF* models) and ST-based predictions (i.e. the *AR+SDSI* and *RF+SDSI* models) highlights that the nonlinear model is able to extract more information

2. In this case, we do not consider the *CAR(local)* and *CRF(local)* models which require NWPs data to perform weather conditioning.

from ST data sources than the linear model. For instance, for a 1-hour lead-time, ST information improves the nRMSE score by $18.1 - 8.7 = 8.0\%$ when considering the RF model, while, we observe an increase of $10.1 - 5.0 = 3.7\%$ with the ARX model.

#### 6.4.3.3 PV production observations + NWPs

**6.4.3.3.1 Explanatory features** This approach explicitly considers the information carried by the NWPs data. The *AR+NWP* model manages to extract relevant information in such a way that it can improve nRMSE scores by up to 22.2% in comparison with the *AR* model for 6-hour lead time. Accuracy improvement due to weather information increases with lead time. Similar conclusions are drawn when comparing the *RF+NWP* model with the *RF* model. Nevertheless, for all considered horizons, the *RF+NWP* model outperforms its linear counterpart.

**6.4.3.3.2 State features** This approach considers the NWPs information as a way to gather PV production measured under similar weather states. As a result, the dynamics are directly carried by production observations. The *CAR(NWP)* model slightly performs better than the *CRF(NWP)* model: on average, a performance increase by +1.64% and +0.68% (in terms of nRMSE and nMAE) is observed in favour of the *CAR(NWP)* model.

To validate the quality of the different models, the residuals are checked for normally distributed, and uncorrelated properties. Figure 6.13 represents the error distribution of the *Persistence*, *AR*, *CAR*, *RF* and *CRF* models. First, the error distribution tends to get wider for all models as the look-ahead time gets longer. Compared with other models, the *Persistence* models tend to have centred and symmetrical distributions. On the other hand, the WHCO process has ambiguous influence on the distribution of the forecast error. For instance, the weather conditioning of the *AR* and *RF* models tends to reduce skewness of errors obtained at short look-ahead times (i.e. $h \leq 180$ min), while an opposite tendency is observed for higher horizons. In addition, the distribution curves of the conditioned models tend to be sharper (e.g. the standard deviation of the forecast error obtained with the *CAR(local)* model is never greater than 10.08% of $P_c$, while the distribution of the *AR* model error can reach 14.82% of $P_c$).

Figure 6.14 represents the Auto-Correlation Function (ACF) of the residuals obtained with the *Persistence*, AR and RF models, as well as the weather conditioned AR and RF models. For all of the models considered, we observe that as the forecast horizon gets higher, so does the residuals correlation. For the 15-min ahead lead time, models possess uncorrelated residuals except for the very first lag. On the contrary, for higher horizons, there are patterns in the residuals: the auto-correlation values are significant for a higher number of lags. Thus, this graph suggests that for these models there is still information left in the residuals, and that better models exist. It is insightful to note

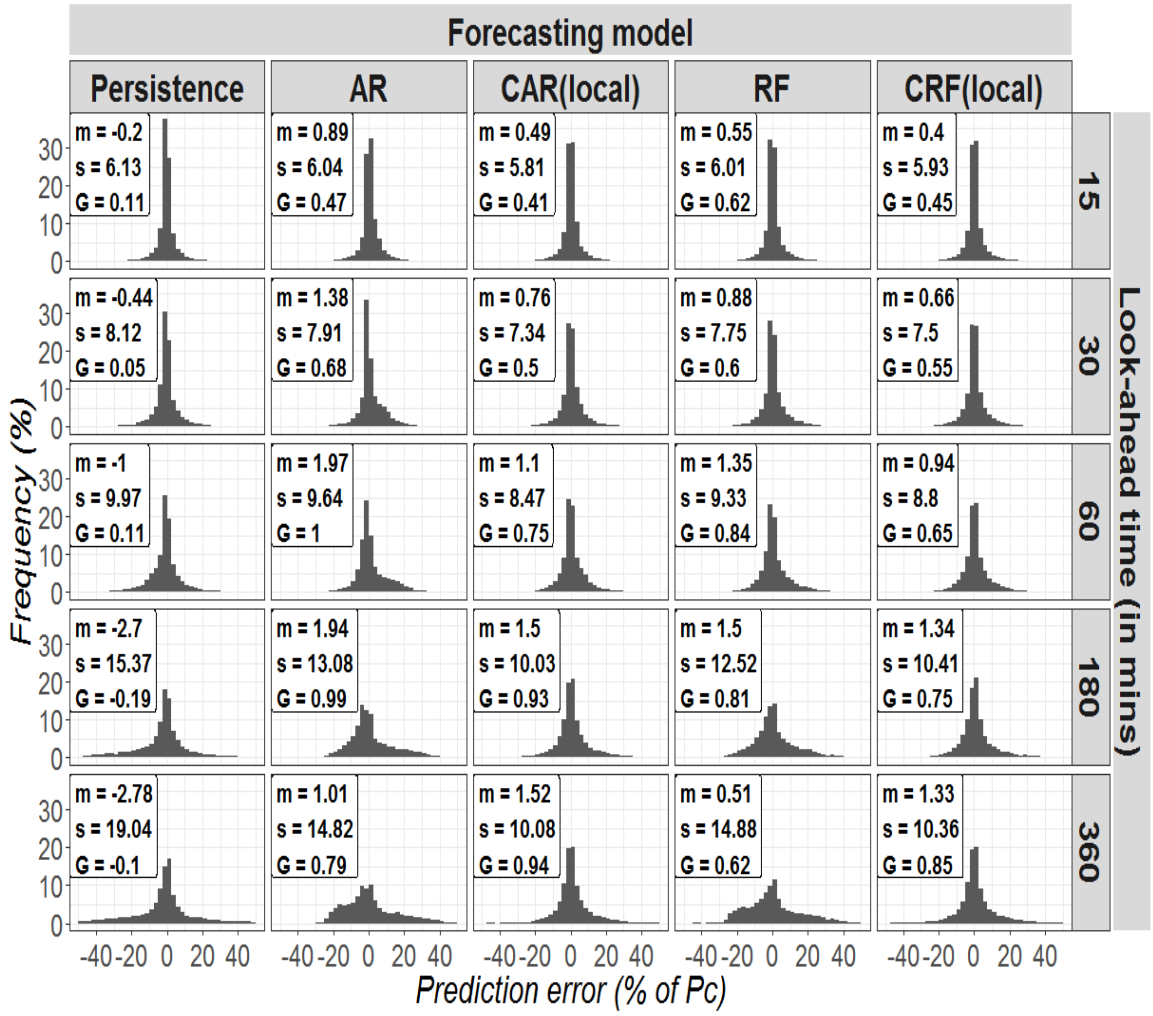



Figure 6.13 – Distribution of normalised prediction errors (with bins representing 2.5% of the rated power) at PV1 for the *Persistence*, *AR*, *CAR(local)*, *RF* and *CRF(local)* models according to the look-ahead times. Values in the upper-left corners represent the mean (m), standard deviation (s) and skewness [a] (G) of the distributions. They are expressed in % of $P_c$.

*a.* Skewness is a measure of symmetry: a negative/positive value indicates that the mean of the data is less/larger than the median, and the data distribution is left/right-skewed.

that, in general, the weather conditioning tends to slightly decrease the value of the auto-correlation. For instance, at a 6-hour lead time, the ACF of the *CRF(local)* model displays a more pronounced exponential decaying that the *RF* model. Thus in this situation, forecast error at a specific time is less correlated to the previous forecasts errors. This may reflect the adaptive capabilities of the model to deal with sudden weather changes.

**6.4.3.3.3 Explanatory and/or state features** In a next step, we focus on the best way to integrate NWPs in a forecasting model. In the case of the linear model, the conditioning approach exhibits higher forecasting performances (i.e. the *CAR(NWP)* model is better than the *AR + NWP* model for both metrics). Based on observations from Section 6.4.2, this is assumed to result from a better integration of features that have a nonlinear correla-

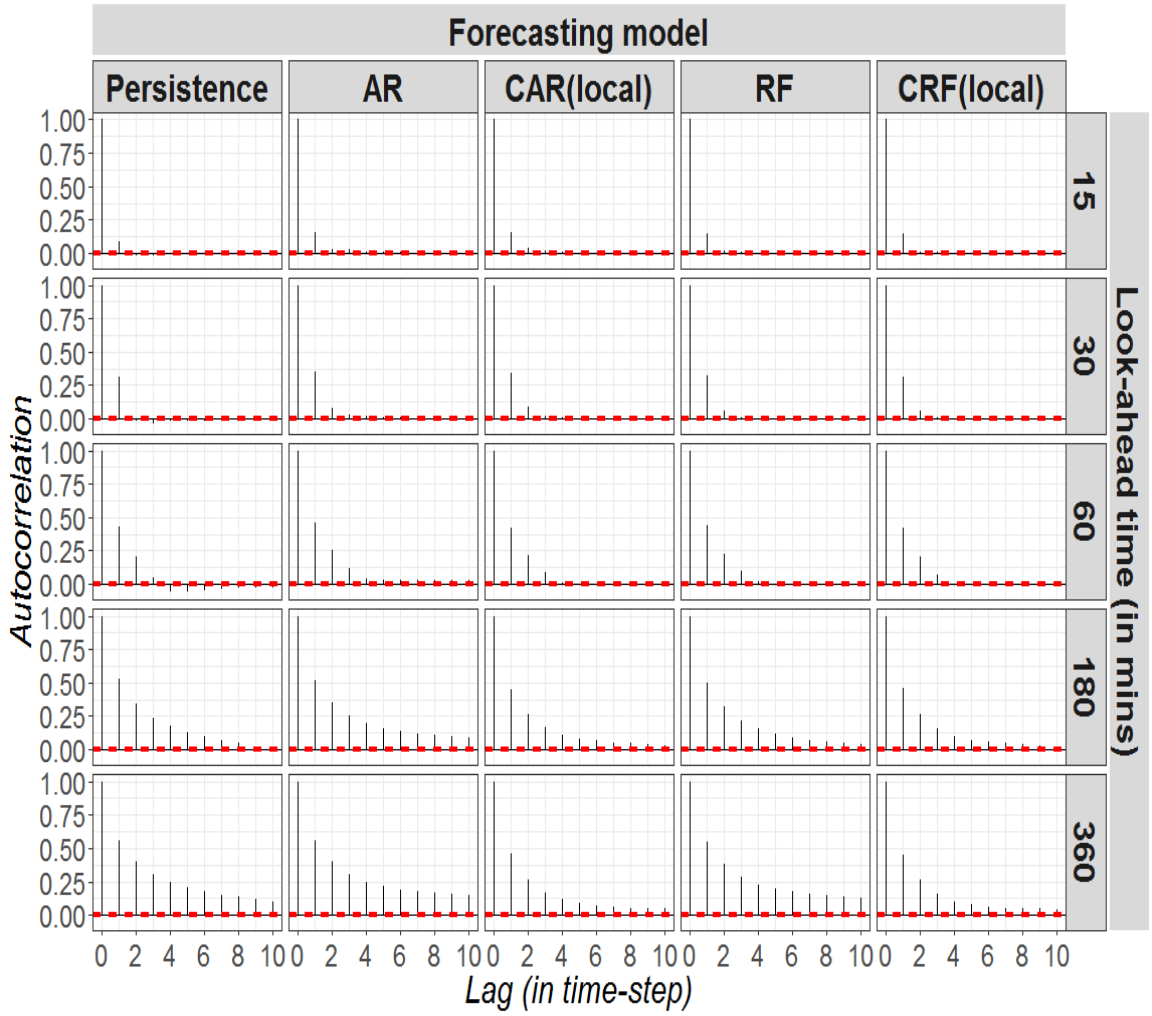


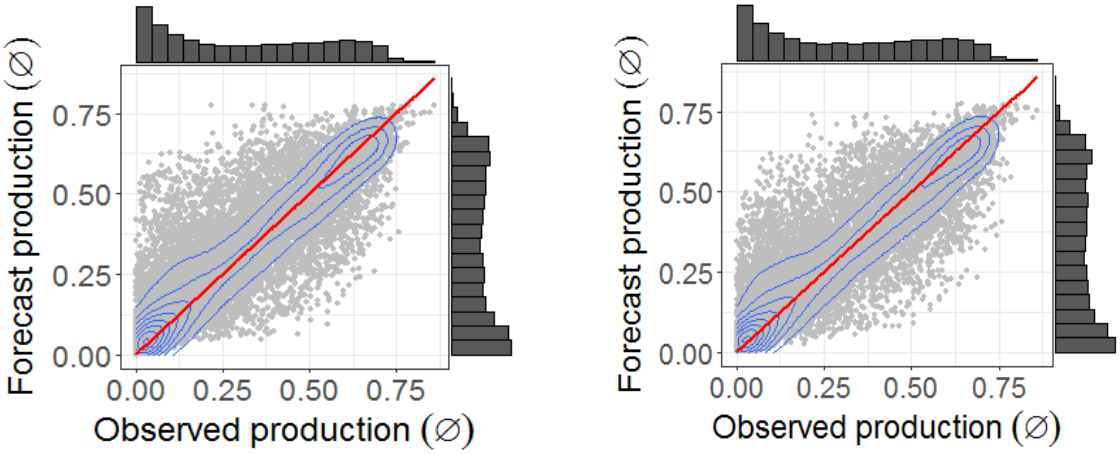

Figure 6.14 – ACF of the time series of errors for some selected horizons obtained with the *Persistence*, *AR* and *RF* models as well as their WHCO forms. The difference between the Clear-Sky Index (CSI) of the forecast and the CSI of the observation is analysed. PV1 is considered. The red dashed lines represent the 95%-confidence bounds.

tion with the production. On the contrary, when dealing with nonlinear forecasting models, it is better to include NWPs as explanatory features. Despite the improved performances due to the conditioning approach, the *CAR(NWP)* model is outperformed by the *RF+NWP* model.

It is possible to further improve the forecasting performance of the *CAR(NWP)*) model by adding NWPs as extra features (which leads to the *CAR(NWP)+NWP* model). This configuration leads to similar performances as those reached by the *RF+NWP* model (i.e. on average, a 0.06% and −0.46% difference is observed between the *CAR(NWP)+NWP* and *RF+NWP* models in terms of nRMSE and nMAE scores). In this respect, the integration of NWPs as explanatory features in the *CRF(NWP)* model slightly decreases its forecasting skills, possibly due to overfitting issues.

As a result, when dealing with production observations and NWPs, the best choice is

either to consider nonlinear models fed with explanatory features or WHCO linear models with weather-based explanatory features. In this regard, Figure 6.15 reveals that both models generate forecasts with similar distribution (only the forecasts generated for the 3-hour lead time are represented, but similar conclusions are reached with other forecast horizons): models tend to overestimate low-production levels. At this point, the choice of the model results mainly from a computational cost and interpretability compromise.




(a) Forecasts obtained with *CAR(local)+NWP* model.

(b) Forecasts obtained with *RF+NWP* model.

Figure 6.15 – Joint and marginal distributions of 3-hour ahead forecasts and production observations at PV1. The contour lines represent the 2D kernel densities, while the red line is the first bisector of the graph. Marginal plots constitute histograms of forecast and observed production.

#### 6.4.3.4 PV production observations + NWPs + SDSI

**6.4.3.4.1 Forecasting performances** When dealing with PV production, NWPs and SDSI inputs, it is obvious that including these data as explanatory features in an *AR* model leads to the worst performances both in terms of nRMSE and nMAE. Once again, the WHCO approach improves significantly the forecasting performances of the linear model. For instance, at a 15-min look-ahead time, the *CAR(local)+SDSI* model surpasses the *AR+NWP+SDSI* model by $11.4-3.2=8.2\%$ and $4.3-(-5.8)=10.1\%$ in terms of nRMSE and nMAE. In addition, we observe that the nRMSE score of the *CAR(local)+SDSI+NWP* model is slightly better than that of the *CAR(local)+SDSI* model.

Conditioning nonlinear models results in a decrease in performances compared to the *RF+NWP+SDSI* model for both metrics.

To conclude, the *CAR(local)+NWPs+SDSI* model appears to be a good option inasmuch as it performs better on very short-term horizons and exhibits similar skill scores to the *RF+NWPs+SDSI* model for high forecast horizons.

Table 6.1 summarises the different findings regarding the choice of the optimal model according to the inputs.

| Inputs | Best configuration |
|---|---|
| PV production | RF |
| PV production+NWP | CAR(NWP)+NWP / RF+NWP |
| PV production+SDSI | RF+SDSI |
| PV production+NWP+SDSI | CAR(NWP)+NWP+SDSI / RF+NWP+SDSI |

Table 6.1 – Summary of the best model configuration (in terms of accuracy criteria) depending on the type of input.

**Research Answer - *Weather conditioning and model family***

On the one hand, the weather conditioning approach is well adapted for linear model in the sense that it naturally improves its capabilities, especially with features that have nonlinear dependencies on the response variable. On the other hand, this approach seems redundant with nonlinear models. In that case, it can induce performances degradation for very short-term horizons compared to a direct integration of exogenous features. Due to its higher computation cost, WHCO is not suitable for nonlinear models. These findings tend to support that weather conditioning is not adapted to nonlinear models, therefore, it should be used carefully in the literature.

**6.4.3.4.2 Model adaptability** Figure 6.16 depicts the Feature Relative Importance (FRI) of the satellite-based information obtained from the regression coefficients of the *CAR(local)+SDSI* model and Equation 6.11. The right-hand graph highlights the annual variability of the FRI of the SDSI dataset, which justifies the use of the WHCO approach as a way to dynamically update the model parameters. This graph also shows that southernmost features, namely points 6 and 8, contribute little compared to westward points located at a similar distance from the park (i.e. points 7 and 10). Points 2, 5 and 9 visually exhibit a seasonal dependence. Surprisingly features 2 and 9, which are in the same direction, show opposed behaviours: the FRI of feature 2 is higher during summertime, while feature 9 is prevailing during wintertime. This may indicate various wind regimes: low wind-speeds occurrence is higher during summertime, while higher wind-speeds are associated with the winter season. In this configuration, the forecasting model has to look further in space to get information regarding incoming clouds. The graph also highlights that westward pixels provide more information.

$$FRI_t^j = \frac{\sum\limits_{\substack{l\in\{-L\ldots0\}\\ s=j}} |\beta_{l,s,t}|}{\sum\limits_{\substack{l\in\{-L\ldots0\}\\ s\in\{1\ldots N_{SDSI}\}}} |\beta_{l,s,t}|} \tag{6.11}$$

| | |
|---|---|
| $j$ | Considered SDSI feature, |
| $N_{SDSI}$ | Number of SDSI features selected (here, $N_{SDSI} = 10$), |
| $L$ | Lag order of the ARX model, |
| $\beta$ | Regression coefficient of the ARX model. |

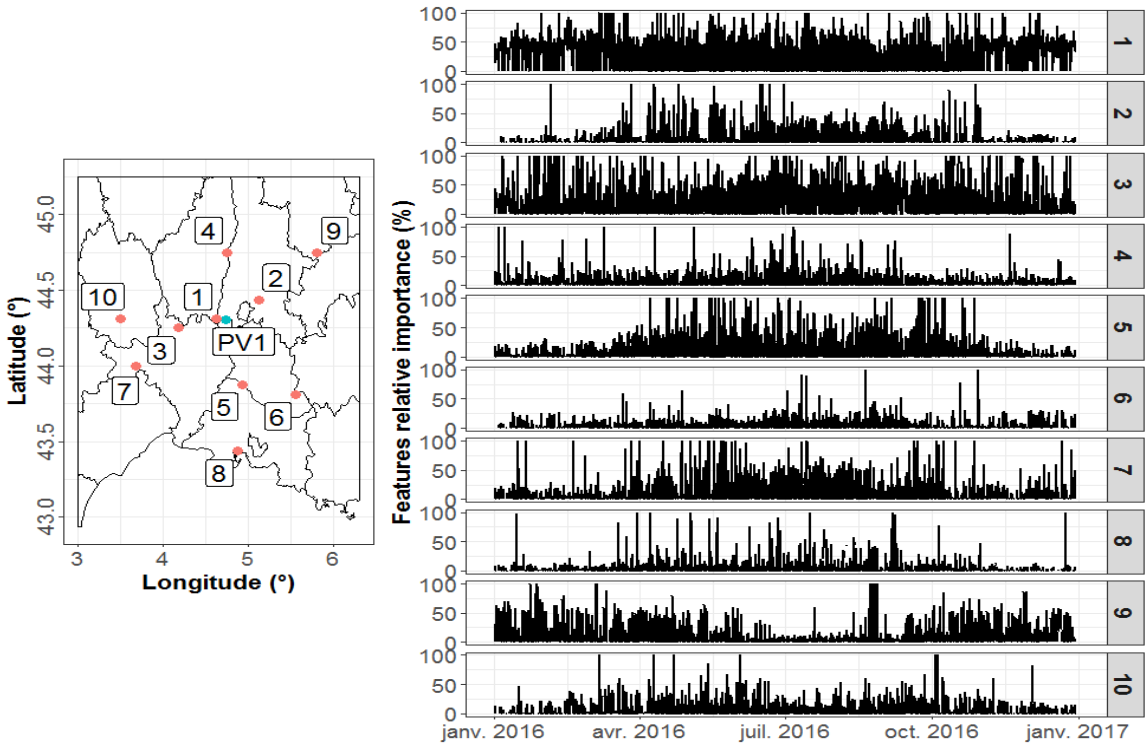



Figure 6.16 – The left-hand figure represents the spatial distribution of selected features from SDSI maps, obtained with the mRMR feature selection approach (Section 5.3.2.1.2) for a 1-hour forecast horizon at site PV1. Features (red points) are ordered by their distances from the park (blue point). The right-hand graph represents the temporal evolution of the FRI obtained with the *CAR(local)+SDSI* model coefficients and Equation 6.11.

**6.4.3.4.3 Coupling of spatio-temporal information with the weather-conditioning approach** In Section 5.3.3, we compared forecasting performances of an ARX model fed with satellite-based information derived from the mRMR feature selection process and from a CNN forecasting architecture. Figure 5.17 of the last chapter clearly depicts the interest of pre-processing satellite-based maps with a Deep Neural Networks (DNN) architecture (i.e. the *AR+SDSI(t+h/CNN)* model outperforms the *AR+SDSI(t-9:t/mRMR)* model up to 15% in nMAE score). Therefore, CNN-derived forecasts are more informative than the set of spatially distributed SDSI observations when injected in the ARX architecture. These

satellite-derived features are now compared in the light of the Conditioned Auto-Regressive (CAR) architecture. Figure 6.17 highlights that the difference in performances between the *CAR(local)+SDSI(t-9:t/mRMR)* and *CAR(local)+SDSI(t+h/CNN)* models is lower than in the case of the *AR+SDSI(t-9:t/mRMR)* and *AR+SDSI(t+h/CNN)* models (Figure 5.17). This slight performance difference between ST models may suggest that the WHCO architecture is be able to emphasise information contained in past production observations, that is similar, and so redundant, to that carried by the CNN-based irradiance forecasts.

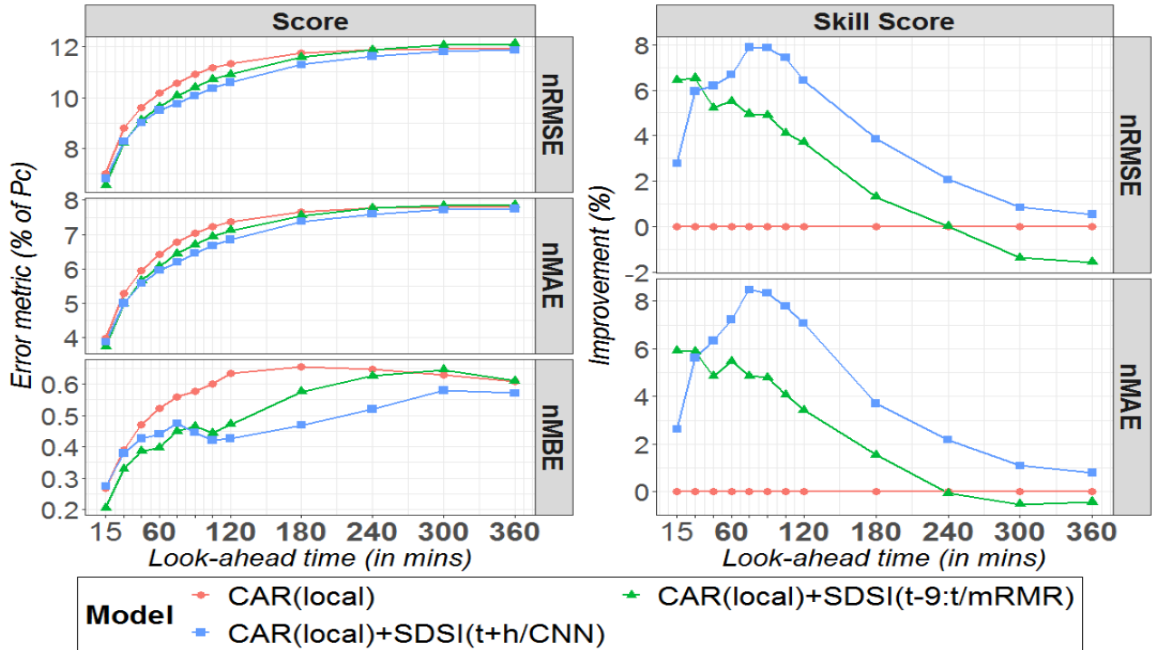



Figure 6.17 – Forecasting performances of the *CAR(local)* models fed with SDSI-based explanatory features. Satellite-based information is extracted via the mRMR feature selection process (i.e. *SDSI(t-9:t/mRMR)*) (Section 5.3.2.1.2), or it is preprocessed to derive forecasts with a CNN forecasting architecture (i.e. *SDSI(t+h/CNN)*) (Section 5.3.2.3). The terminology used is derived from Section 5.3.3.

In addition, we observe that the coupling of WHCO approach and ST information is well adapted to very short-term horizons. Indeed, the *CAR(local)+SDSI(t-9:t/mRMR)* model exhibits improved scores up to +6% in nRMSE and nMAE compared to its temporal counterpart (i.e. the *CAR(local)* model) for a 15-min lead time, while the ST-induced improvement is not significant when considering the ARX model (Figure 5.17). Negative skill scores of the *CAR(local)+SDSI(t-9:t/mRMR)* for higher forecast horizons are assumed to result from overfitting issues. In such a case, it could be relevant to draw inspiration from Section 5.2.3 to impose some constraints over the number of temporal lags. We observe that for horizons greater than 30-min ahead, the model based on the CNN-derived feature has better scores than the model based on the set of selected features. This assumes that accuracy improvement from ST inputs is still possible when considering selected features with the mRMR algorithm. At least two options are conceivable: we can either increase the number of selected features at the risk of overfitting the model or consider a WHCO

approach that respect ST dependencies. This last option is investigated in the following section. So far, a local characterisation of the atmosphere has been considered. Such an approach only takes into account the temporal evolution of weather parameters at the site position. In the next step, the idea is to widen the spatial window on which the analog research is performed to identify situations that evolve likely both in time and space.

## 6.5 Synoptic weather information

The present section compares the benefits provided by WHCO based on spot- or large-scale weather predictions. The influence of the coupling between weather information scale and ST inputs is then investigated. Only forecasts obtained with the ARX model are considered, as the previous section highlighted that WHCO approach is not fitted for the RF model (degradation in the forecast accuracy for very short-term horizons and higher modelling complexity compared to the direct use of the RF model).

### 6.5.1 Temporal paradigm

As a first step, we focus on assessing the forecasting performances of the WHCO models that only use temporal measurements, in other words, SDSI observations are left aside.

In the case of WHCO with gridded features, the best forecasting performances are achieved with a higher number of analog situations compared to spot-NWPs conditioning (Figure 6.18). This need for a higher number of training observations may be explained by the variability of the training set itself, which is higher in the case of a conditioning based on geopotential fields due to the nature of the predictor and the extent of the spatial window considered. The WHCO based on spot-NWPs focuses on parameters observed at the site location that directly impact PV production, while its counterpart based on gridded data considers features impacting wind and cloud generation on a much larger scale. Therefore, local weather situations may vary significantly even though the associated geopotential fields are rather close (e.g. a high-pressure area is associated with clear skies, yet local cloud structures may be present).

The fact that synoptic WHCO provides a set of observations that share similar large-scale dynamics but potentially different local states may explain the poor performances of the *CAR(synoptic)* model, observed in Figure 6.19, w.r.t. the *CAR(local)* model. Within a set of analog situations provided by the synoptic WHCO approach, local weather states are too heterogeneous to allow the establishment of an accurate forecasting model.

### 6.5.2 Spatio-temporal paradigm

Contrary to the spot conditioning, which only considers temporal evolution of features, synoptic conditioning selects a set of situations that follow similar ST trends (e.g. same

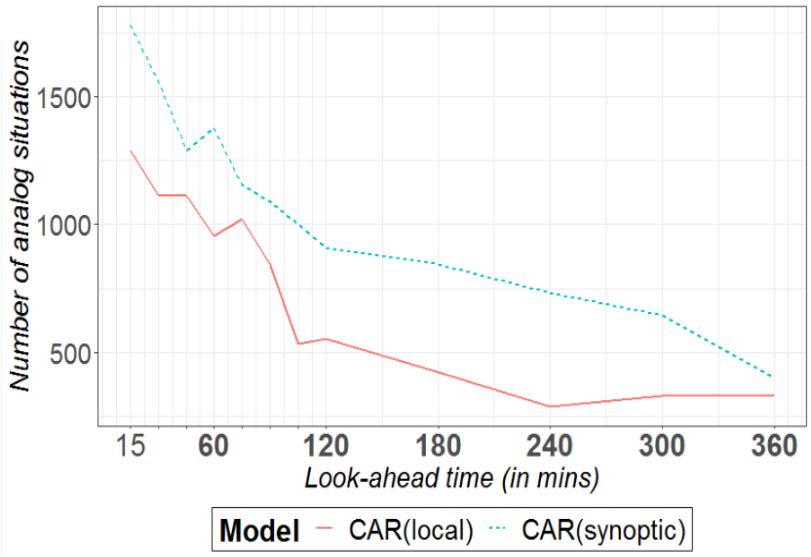



Figure 6.18 – Averaged number of analog situations determined with the optimisation framework for the 9 PV farms under study and that considers spot or gridded NWPs.

wind direction). In this section, the coupling of the synoptic WHCO architecture with ST inputs is investigated.

Therefore, we focus on performance improvements resulting from the consideration of SDSI features w.r.t. a configuration based only on temporal observations (i.e. production observed at the site location). The left-hand graph in Figure 6.20 represents the ST gain in accuracy obtained with a local conditioning, while the right-hand graph represents the same gain considering a synoptic conditioning. The trend of the performance gain obtained from ST information is similar between both approaches: a maximum is reached for very short-term horizons (typically $h \leq 90$ min) and it gradually decreases as the forecast horizon increases. A visual comparison highlights that the gain obtained with SDSI is slightly higher in the case of the synoptic conditioning. It is difficult to determine whether this improvement results from a better integration of the ST information (e.g. the selection of SDSI features is in line with the wind regime of the analog situations) or from the lower performances achieved by the *CAR(synoptic)* model which lets more room for improvement. Be that as it may, the performances developed by the *CAR(synoptic)+SDSI* model do not beat its counterpart based on local conditioning (Figure 6.21). It could have been interesting to combine both local and synoptic weather parameters within the analogy score (Equation (6.2)) to obtain situations that are analog at the site position and its vicinity. A similar idea is implemented in [67] that considers parameters at the site location as well as a set of SDSI features located in the neighbourhood.

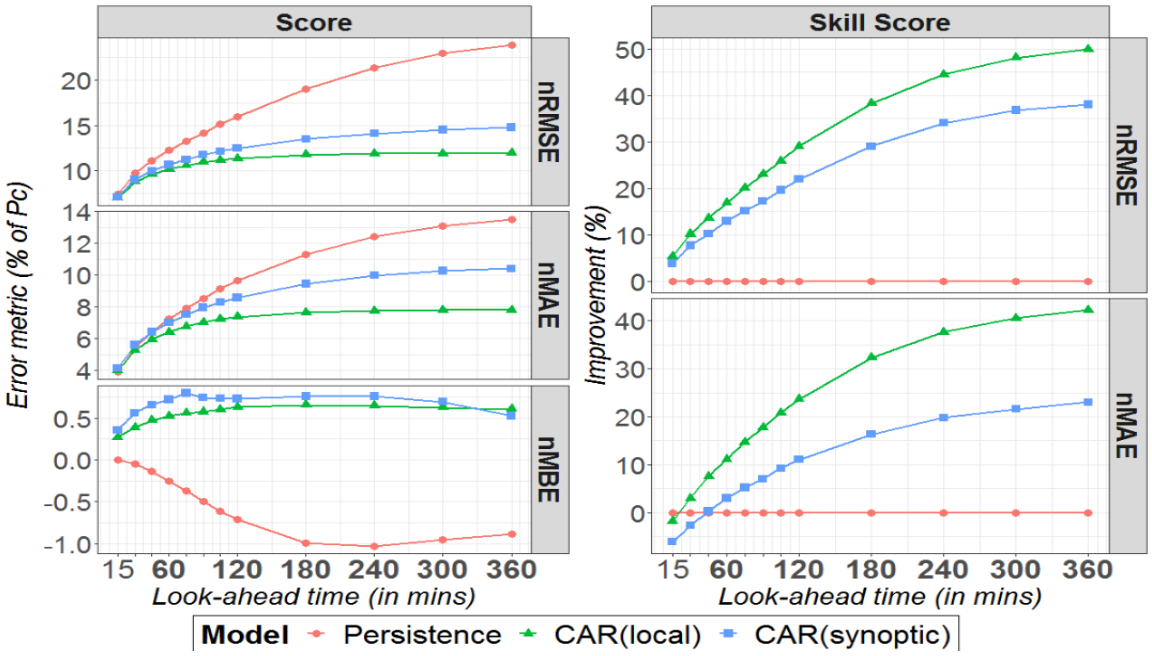
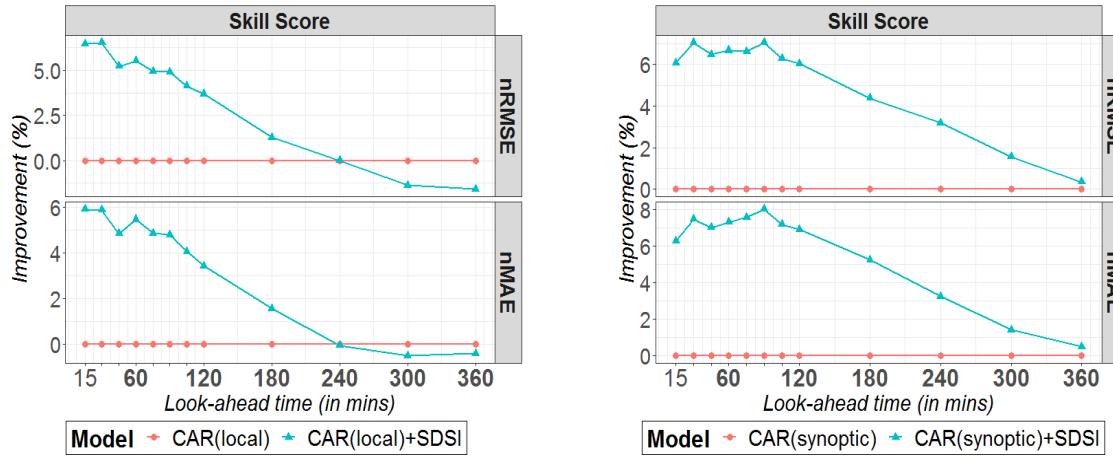



Figure 6.19 – The *CAR* model conditioned either with local or synoptic features and fed with PV production observations.




(a) Influence of ST inputs on the forecasting performances of the locally conditioned model.

(b) Influence of ST inputs on the forecasting performances of the synoptically conditioned model.

Figure 6.20 – Skill scores of ST models compared with temporal models.

**Research Answer - *Gridded data***

The use of gridded data, and more precisely geopotential fields, do not improve forecast accuracy, neither in the scope of temporal-based forecasts nor with the use of ST features. This may result from a too high variability within the candidate situations sets (and so, from a too restricted historical archive), which prevents the derivation of accurate models.

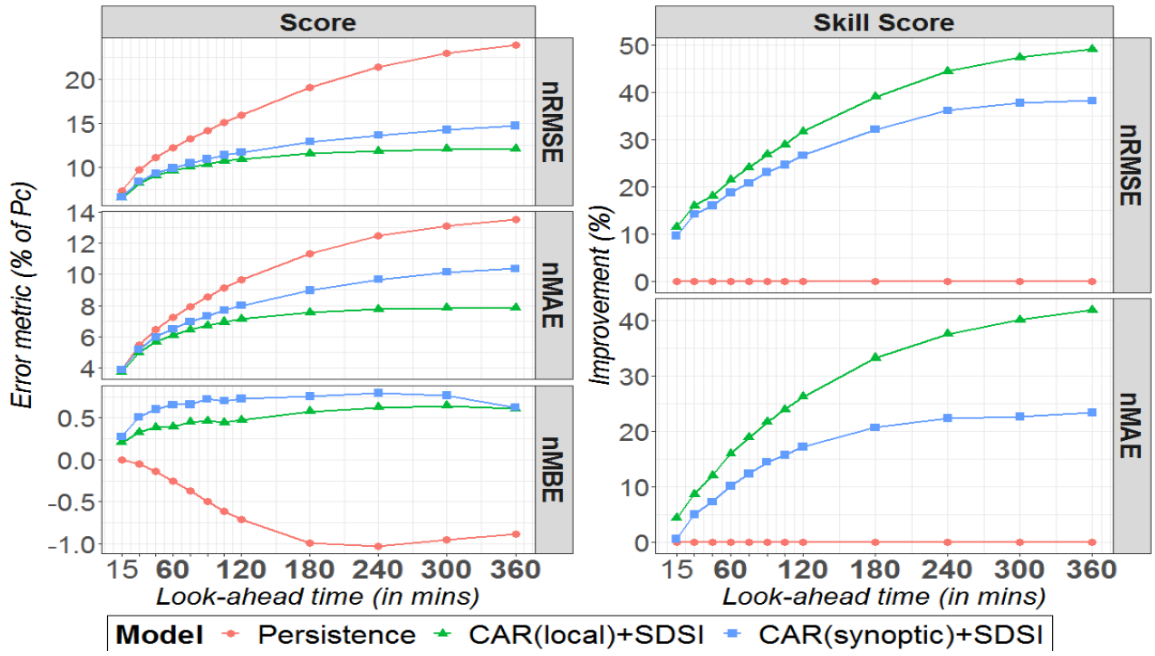



Figure 6.21 – The *CAR* model conditioned either with local or synoptic features and fed with PV production observations and SDSI observations.

## 6.6 Probabilistic forecasting

A forecast is inherently uncertain. To be able to quantify this uncertainty can be valuable in a context of decision-making [273]. In contrast with deterministic (or point) forecasting, probabilistic forecasting provides forecasters with additional information regarding the uncertainty of a forecast. This uncertainty can take various forms, for instance, we find ensemble forecasts especially in the field of weather forecasting, where a numerical model is run with perturbed sets of parameter schemes and/or initial conditions to issue different trajectories. We also find Prediction Interval (PI) that estimates the interval in which a future observation is expected to fall, with a certain probability.

These probabilistic forecasts can be generated by dedicated models, or through generic techniques to transform point forecasts into probabilistic ones. For instance, a generic approach consists in bootstraping[3] PI from empirical errors obtained during the training step. More recently, the Level Set Forecaster (LSF) technique has been presented in [88, 274]. The main idea is to gather instances of explanatory features from the training data associated with close predictions, and then to derive predicted distributions from bins composed of observations of the response feature associated with these selected inputs. In other words, this approach identifies a set of training examples, $X_t$, that are mapped to the same (or close) point forecast value $\hat{y}_{t+h} = f_{root}(X_t)$ from the testing set. Their corresponding true target values (i.e. $y_{t+h}$) from the training set are then collected within a bin. The bin associated with a predicted value $\hat{y}_{t+h}$ is then selected to produce forecasts intervals by picking the

3. Bootstrapping is a type of resampling technique with replacement.

$q^{th}$ quantile of the true values in the bin. Bootstrapping techniques seem to be first applied in the PVPF field in [226] (2015) but very few details were given regarding the underlying process. Later in 2020, [227] proposed to derive prediction intervals by using bootstrap or Quantile Regression (QR). The results demonstrate that the bootstrap approach leads to the best performances. Confidence intervals for spot wind generation forecasting are assessed from forecast error re-sampling in [275]. In the present context, such approaches can be implemented while taking advantage of the analog-based structure developed throughout this chapter. For example, we may contemplate using the bootstrap method to derive confidence intervals through the set of past analog PV production observations by selecting randomly and with replacement $N$ values out of the analog sample. Then, this new sample is sorted in ascending order and the 2.5% lowest and 97.5% highest values of that set are selected. These three steps are repeated a large number of times to get a good idea of the population. Such an approach should provides sharp intervals for analog subsets with low variability. Nonetheless, due to time-constraints, we chose to make good use of the proposed modular architecture by replacing deterministic root models with probabilistic ones. *This section must be viewed as a presentation of some preliminary works on the weather conditioning of probabilistic forecasts. As a result, further developments are needed to consolidate the results.*

### 6.6.1 Probabilistic models

We replace the ARX and RF models with their probabilistic counterparts in a way, that is to say the QR [4] [276] and QRF [5] [277] models. The aim of these models is to approximate the conditional distribution of the random variable (in this case, PV production) by means of quantiles. The $\tau^{th}$ quantile of the random variable $Y_t$ is defined as $P(Y_t < x) = \tau$ where $\tau \in [0, 1]$. A unique quantile provides only limited information regarding forecast uncertainties, that is why we consider the following quantiles: $\tau \in [0.1, 0.2, ..., 0.8, 0.9]$. The QR and QRF models estimates the $\tau^{th}$ conditional quantile function $q_{t+h|t}(\tau)$.

### 6.6.2 Diagnostic analysis

In this section, we limit our study to the most common criteria of probabilistic performances assessment, namely the reliability and the sharpness [278].

#### 6.6.2.1 Reliability diagram

Reliability or calibration concept describes the ability of probabilistic forecasts to match the observation frequencies. Reliability diagrams are a graphic tool that makes it possible to

4. This model expresses the conditional quantiles of the response feature as a linear function of the explanatory variables.

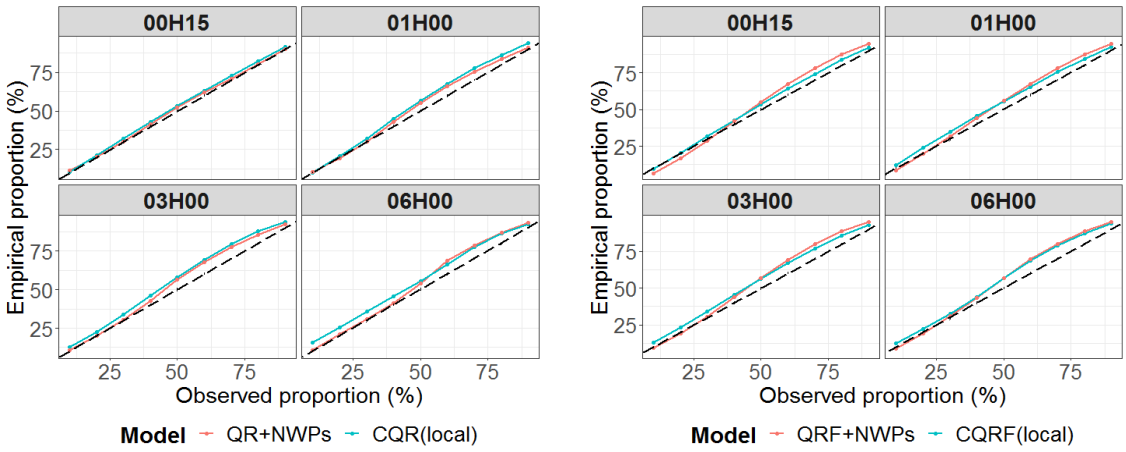
5. Compared to the RF model which generates forecasts by averaging values in the leaf nodes, the QRF model derives probabilistic predictions from these values.

verify whether the proportion of data predicted by the quantile $\hat{q}^{\tau}_{t+h|t}$ of level $\tau$ is equal to the associated proportion of actual data observed under this quantile. Ideally, the proportions of each quantity should be close. Therefore, the closest to the diagonal, the better. The empirical level of quantile forecasts is obtained through the following steps [279]. First, the indicator function $\xi^{\tau}_{t,h}$ defined by Equation 6.12 is computed. It represents the series of *hits* (i.e. the production outcome lies below the quantile forecast) and *misses* of the evaluation set. The empirical level $a^{\tau}_{h}$ of these quantile forecasts is given by the mean of $\xi^{\tau}_{t,h}$ over the evaluation set.

$$\xi^{\tau}_{t,h} = 1_{y_{t+h} < \hat{q}^{\tau}_{t+h|t}} = \begin{cases} 1 & \text{if} y_{t+h} < \hat{q}^{\tau}_{t+h|t} \\ 0 & \text{otherwise} \end{cases} \tag{6.12}$$

$\hat{q}^{\tau}_{t+h|t}$ Quantile forecast issued at time $t$ for lead time $t+h$,
$y_{t+h}$ PV production outcome.

Figure 6.22 gathers the reliability diagrams of the QR and QRF models as well as their WHCO forms. We observe that, in general, the different forecasts seem reliable due to their close proximity with the diagonal. A closer analysis reveals that the model reliabilities tend to decrease as the forecast horizon increases. The *CQRF(local)* model tends to exhibits slightly better reliability properties than the *QRF+NWPs* model. On the contrary, the *QR+NWPs* model is more reliable than the *CQR(local)* model.




(a) Reliability diagram of the QR model fed with exogenous inputs or conditioned to the local weather.

(b) Reliability diagram of the QRF model fed with exogenous inputs or conditioned to the local weather.

Figure 6.22 – Reliability diagrams of forecasts at PV1. The dashed line represents forecasts with perfect reliability. The headers of the sub-graphs represent the forecast horizons.

#### 6.6.2.2 Rank histogram

Rank histograms (sometimes called Talagrand diagrams) are another graphical way to assess the reliability of an ensemble forecast compared to a set of observations [6]. Thus, it aims at checking the statistical consistency of the ensemble forecasts (i.e. that the predicted probabilities agree statistically with the observed frequencies).

Rank histograms are built as follows. For each instance $n \in (1, \cdots, N_t)$ ($N_t$, the total number of observations from the testing set), the $M$ ensemble forecasts and the associated observation are ranked together in an ascending order. The rank of the observation within the group of $M+1$ member is computed [7]. A histogram is then derived from these ranks.

The main assumption behind rank histogram is that an observation is statistically another member of the ensemble forecasts. In this context, the observation is equally likely to fall between any two members. This leads to a flat diagram (i.e. a uniform distribution): the observations are indistinguishable from any member of the ensemble and the ensemble forecasts are said to be reliable. If the histogram is $\cup$-shaped (or $\cap$-shaped), then the spread of the ensemble forecasts is too small (or to large): many observations fall at the tails of the ensemble (or near its centre). A tilt or asymmetric histogram suggests that too many observations fall outside (below or above) the ensemble members, which is typical of a biased forecast.

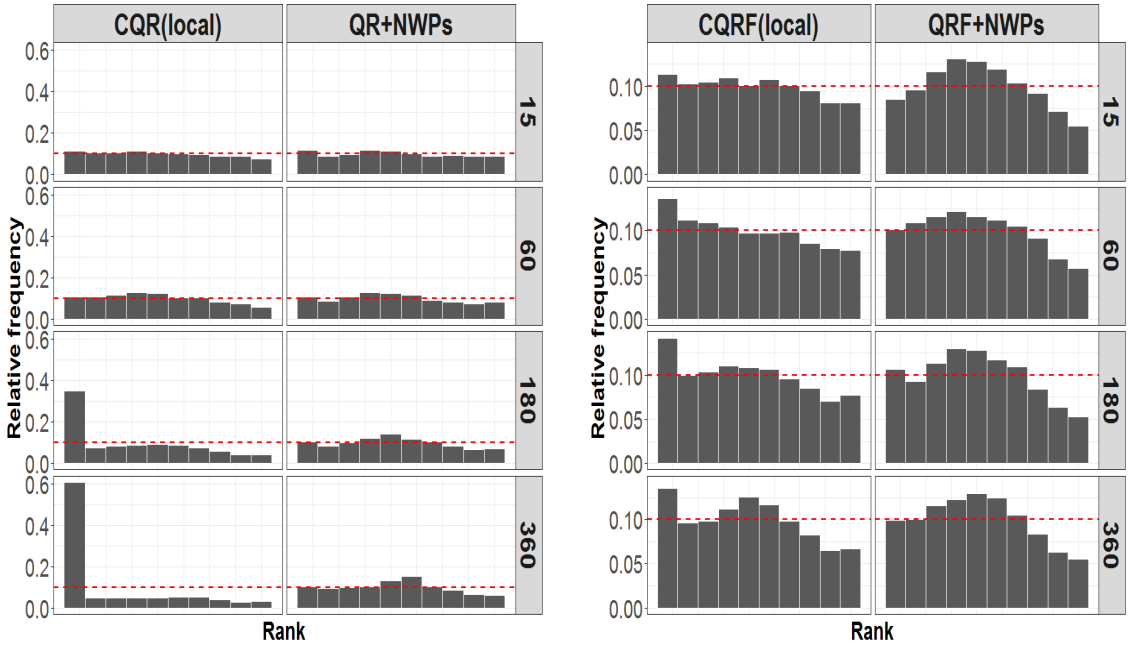
Figure 6.23 represents the Talagrand diagram of QR and QRF models fed with NWPs information either as explanatory or state features. At first glance, we observe that for the great majority of models, the different ranks are close to the consistency band (i.e. the red dashed line) for all considered horizons. The only exceptions are the *CQR(local)* model associated with the 180-min and 360-min ahead forecast horizons, which demonstrate high relative frequencies for the first rank. Overall, histograms are characterised by a combination of two distinct shapes; namely a prevalence of the first rank and a convex shape. This indicates that associated ensemble forecasts tend to be over-dispersed and biased (i.e. production associated with clear-sky states is under-estimated, while overcast states lead to an over-estimation of production). Models considering weather predictions as exogenous inputs tend to express a convex shape, this phenomenon is corrected by the weather conditioning approach at least for the 15-min ahead horizon (i.e. the *CQR(local)*- and *CQRF(local)*-derived forecasts possess a flat rank distribution). Therefore, the analysis of the rank histograms highlights that ensemble forecasts possess some reliability flaws. The latter may be corrected by post-processing ensemble forecasts techniques such as Ensemble Model Output Statistics (EMOS) [8] [280]. In the remainder of this section, we neglect these reliability flaws due to the fact that histograms are dominated by a high number of rank

6. This graphical tool is often preferred by the meteorologist community because the reliability diagram may be difficult to visually assess.

7. For instance, for the set of ordered predictions $(0.12, 0.43, 0.51, 0.55, 0.62, 0.65, 0.70, 0.79, 0.81)$, the rank of the observation (0.6) is 5, because the observation lies at the $5^{\text{th}}$ position between 0.55 and 0.62.

8. Technique based on multiple linear regression that addresses both forecast bias and under-dispersion.

distributions close to the consistency band. However, we draw the reader's attention to the fact that special attention should be paid to conclusions regarding the *CQR(local)* model. This leads to further investigations about the sharpness of the forecasts.




(a) Rank histograms of the QR model fed with exogenous inputs or conditioned to the local weather.

(b) Rank histograms of the QRF model fed with exogenous inputs or conditioned to the local weather.

Figure 6.23 – Verification rank histograms of PV1 considering the *CQR(local)*, *QR+NWPs*, *CQRF(local)*, and *QRF+NWPs*, models at 15-, 60-, 180-, and 360-min lead times. The red dashed line represents the theoretical relative frequency for a uniform distribution (here 1/10).

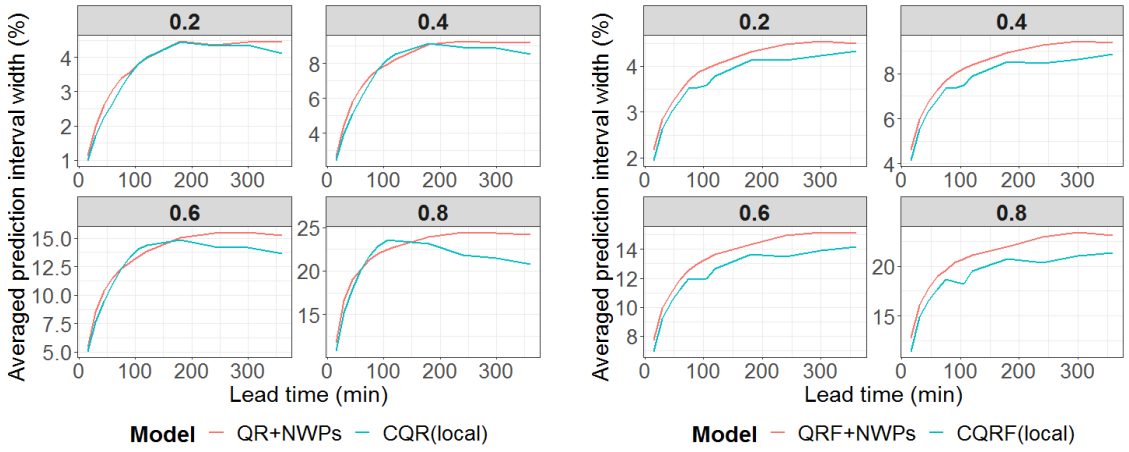
#### 6.6.2.3 Sharpness assessment

Sharpness evaluates the concentration of predictive distributions, as such it constitutes a complementary analysis tool to the reliability but does not provide any indication regarding the quality of the forecasts. In simple words, the sharpness evaluates how tight the predictive densities are, regardless of their forecasting abilities. Sharpness is determined from the average width of centred PI. The width of a given PI, $\hat{I}^{\tau}_{t+h|t}$ is the distance between its two bounds for a given quantile level and horizon, $h$ (Equation 6.13). The sharpness of these interval forecasts is the average of $\delta^{\beta}_{t,h}$ over the evaluation period. Therefore, given that the forecasts are reliable, the objective of probabilistic forecasts is to maximise the sharpness, and so to minimise the width of interval forecasts.

$$\delta^{\beta}_{t,h} = \hat{q}^{1-\frac{\tau}{2}}_{t+h|t} - \hat{q}^{\frac{\tau}{2}}_{t+h|t} \tag{6.13}$$

$\beta$ Nominal coverage rate ($\beta = 1 - \tau$).

Figure 6.24 represents the sharpness evaluation of forecasts issued by the QR and QRF models as well as their WHCO forms. On the one hand, concerning the *QR+NWP* and *CQR(Local)* models, we observe that for nominal coverage rates of 20% and 40%, both models exhibit similar sharpness for all considered horizons. However, for nominal coverage rates of 60% and 80%, the *CQR(Local)* model is sharper for lead times greater than 180-min. On the other hand, the conditioned version of the QRF model is sharper for all horizons and nominal coverage rates under study. As a result, the WHCO approach has a positive impact on the model sharpness, whatever the model family (linear or nonlinear).




(a) Sharpness evaluation of the QR model fed with exogenous inputs or conditioned to the local weather.

(b) Sharpness evaluation of the QRF model fed with exogenous inputs or conditioned to the local weather.

Figure 6.24 – Sharpness evaluation of forecasts at PV1 as a function of the forecast horizon. The headers of the sub-graphs represent the nominal coverage rates.

#### 6.6.2.4 Continuous ranked probability score

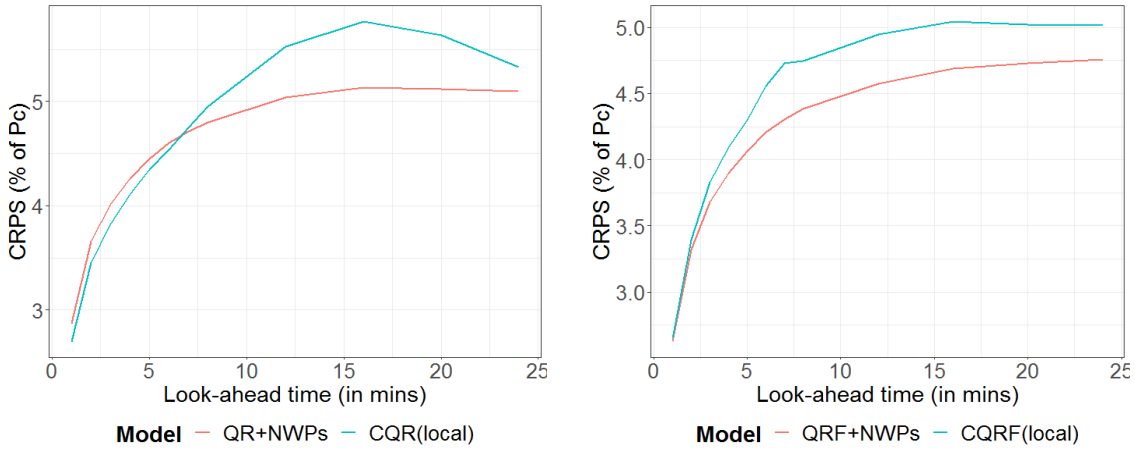
The overall performances of the models are evaluated with the Continuous Ranked Probability Score (CRPS) [281–283]. This score, which is dedicated to probabilistic forecasts, provides summary measures of the forecast quality. It is defined by Equation 6.14 for a given Cumulative Distribution Function (CDF), $F_{t+h|t}$, and its observations $y_{t+h}$. The CRPS is negatively oriented (i.e. the smaller the score, the better), and is similar to the Mean Absolute Error (MAE) when applied to point forecasts.

$$CRPS_{t,h} = \int_{-\infty}^{\infty} \left( \hat{F}_{t+h|t}(x) - H(x - y_{t+h}) \right)^2 dx \tag{6.14}$$

$H(x)$ Heaviside function ($H(x) = 0 \forall x < 0$ and $H(x) = 1$ otherwise).

Figure 6.25 shows the CRPS of the forecasts from the QR and QRF models as well as their WHCO forms. First, a visual comparison between Figure 6.25a and Figure 6.25b

highlights that the *QRF+NWPs* model performs slightly better than other models. In general, models based on the extra-features mode outperform WHCO models, except for forecast horizons lower than 2-hour ahead in the case of the QR model.




(a) CRPS of the QR model fed with exogenous inputs or conditioned to the local weather.

(b) CRPS of the QRF model fed with exogenous inputs or conditioned to the local weather.

Figure 6.25 – CRPS of forecasts at PV1 as a function of the forecast horizon.

**Research Answer - *Probabilistic forecasts***

Weather conditioning tends to positively impact the sharpness of forecasts compared to predictions produced with a straightforward injection of the explanatory features in the model. However, the overall performances of conditioned models are slightly lower than those of models fed with weather variables as explanatory features. Even though differences of performances are slight, the computational cost induced by weather conditioning does not justify a deeper analyse.

## 6.7 Comparison between analog- and cluster-based conditioning

Models can be conditioned to the weather situation at least through two approaches: either by (1) fitting models on clusters of production data associated with similar weather states, and by (2) adopting a dynamical approach that updates models by fitting them on situations that are analog to the situation to predict (this approach is investigated throughout this chapter).

Due to time limitations, the cluster-based conditioning has not been investigated in depth, but preliminary work is proposed in Appendix C.3. The methodology is presented as well as performances comparison between the analog- and cluster-based conditioning. The latter turns out to be significantly less time-consuming (Table A.1), while showing similar

forecasting accuracy when considering the SSRD as a state feature and the AR model. Be that as it may, further works are needed to characterise in detail the influence of the features and model properties on the cluster-based conditioning approach.

## 6.8 Conclusions

In this chapter, we introduced the concept of WHCO based on the analogy principle. This approach is materialised through the development of a generic methodology to integrate weather information (or other features) into PV forecasting models either as explanatory or state features. Within the scope of this thesis, this methodology is applied to short-term PV generation forecasting, but nothing prevents its extension to higher forecast horizons or its use in the wind generation field. The WHCO approach appears as a simple and intuitive method to inject physics-based information in statistical regression models and to derive expert models. This former point goes in the direction of a better interpretability of models.

Performances of the WHCO-based approach depend on the number of analogs used during the model fitting. In essence, this approach aims at reducing the training set to retain only observations with similar characteristics, hence the threat to feed the model with too little data. A grid search approach tackles this issue and shows that the number of analogs leading to the best accuracy depends on the forecast horizon as well as the number of features.

On the one hand, performances analysis highlights that the WHCO approach is well suited for linear models inasmuch as it gets the best from features that have a nonlinear dependence on the response variable. In the case of linear-dependent features, their simultaneous integration as explanatory and state features provides better scores than distinct integrations. On the other hand, when applied to a nonlinear model, WHCO tends to degrade the very short-term forecasting performances and to provide similar scores compared with a straightforward integration of the features. Ultimately, similar forecasting performances are reached when considering either the *CAR(local)+NWPs* model or the *RF+NWPs* model.

In the literature, spot NWPs are typically used to characterise the weather situation at the site location. Such an approach only focuses on the temporal dynamics of the weather parameters and obliterate the ST dynamics. With an ambition to improve the integration of ST features in regression models, we explored the use of geopotential fields as a state feature. However, this approach did not turn out fruitful in terms of accuracy improvement. This may result from an excessively high variability within candidate situations sets (i.e. a too restricted historical archive), which prevents the derivation of accurate models.

A preliminary investigation, which considers the probabilistic versions of the AR and RF models, points out that weather conditioning improves sharpness of the forecasts. Nonetheless, the best overall performances are reached by models fed with NWPs model outputs as

explanatory features.

By way of conclusion, the comparison of the different ways of integrating information in regression models made it possible to derive general guidelines for forecasters.

## 6.9 Résumé en Français

Jusqu'à présent, nous avons considéré une calibration statique des modèles dans la mesure où les paramètres internes de ces derniers sont déterminés lors de la phase d'entraînement puis gardés tels quels par la suite. Ainsi, le calage du modèle reflète les observations les plus représentées dans les données d'entraînement. Ce type d'approche peut s'avérer dommageable notamment dans un contexte spatio-temporel où uniquement les variables localisées selon les axes des vents dominants seront retenues. Pour pallier ce problème, nous faisons le choix d'adopter une approche dynamique permettant la mise à jour des paramètres du modèle selon la situation météorologique rencontrée. Ceci conduit donc à un modèle adaptatif.

### Définition de la méthodologie de conditionnement

L'objectif premier de ce chapitre réside en l'introduction du concept de conditionnement par la situation météorologique et en la définition de ces fondements mathématiques. Typiquement dans la littérature, deux options sont étudiées pour générer des modèles dédiés à des situations météorologiques spécifiques.

La première option consiste à regrouper les données d'observations selon certains critères afin d'obtenir des groupes de données représentant des situations atmosphériques similaires (e.g. des observations ensoleillées, pluvieuses, ou nuageuses), puis à caler un modèle pour chaque groupe. Cette approche est détaillée à l'Annexe C.3.

Dans ce chapitre, nous considérons une approche dérivée des k plus proches voisins qui consiste à ré-entrainer un même modèle sur les N observations les plus similaires à la situation à prévoir. Les différentes étapes de modélisation se déclinent de la manière suivante. Tout d'abord, trois jeux de données sont construits : (1) une archive contenant les prévisions numériques du temps (ce jeu de données caractérise la situation météorologique), (2) l'historique de production du site, et (3) une base de données regroupant les variables explicatives du modèle. Ensuite, un critère d'analogie est utilisé afin de quantifier le degré de similarité entre la situation cible (i.e. la prévision météorologique à l'instant $t + h$) et les situations candidates (i.e. les prévisions météorologiques antérieures générées pour le même horizon temporel). Ces situations sont ensuite classées selon leur degré de similarité. Les $N$ situations les plus similaires sont alors retenues pour constituer le sous-ensemble de situations analogues. Les observations de la production PV associées à ce sous-ensemble ainsi que leurs variables explicatives sont sélectionnées afin d'entraîner le modèle de prévision, alors que les dernières observations des variables explicatives sont utilisées pour générer la

prévision de la production PV à l'instant $t + h$.

D'un point de vue mathématique, cette approche peut être vue comme une régression locale. Ainsi, le conditionnement à la situation météorologique permet de non-linéariser un modèle de régression. Dans ce cas, il est légitime de se demander quelle est l'influence de ce type d'approche sur des modèles non linéaires. Dans la littérature, il est courant de considérer un nombre fixe d'analogues quel que soit l'horizon considéré. Pourtant, on s'attend à ce que le nombre d'observations nécessaires pour dériver une relation statistique stable varie selon la variabilité de la situation météorologique. Nous effectuons donc une recherche de grille afin de sélectionner le nombre optimal d'analogues nécessaires au calage de notre modèle.

### Variables d'état

Outre le critère d'analogie utilisé pour comparer les états de l'atmosphère, une attention toute particulière doit être portée aux variables les décrivant. Celles-ci doivent refléter avec fidélité le processus étudié, qui dans notre cas concerne les variations de la production PV.

Les variations de production sont directement liées aux variations de l'irradiance, elles-mêmes tributaires des perturbations atmosphériques (e.g. formations nuageuses, turbidité de l'atmosphère) et également de la température ambiante. Donc naturellement nous considérons les prévisions de l'irradiance, de la température et de la couverture nuageuse totale au niveau du site d'intérêt. Le format de ces données permet de trouver des situations évoluant de la même manière dans le temps, mais ne garantit aucunement une quelconque similarité parmi les schémas spatiaux des analogues retenus. Ainsi, nous considérons également le champs géopotentiel, une variable 2D couramment utilisée pour prédire les précipitations, et qui démontre une forte corrélation avec les déplacements de masse d'air.

### Principaux résultats obtenus à partir de modèles déterministes

Dans le contexte de variables localisées, nous mettons en évidence que le conditionnement météorologique est bien adapté aux modèles linéaires et permet une meilleure intégration des variables non-linéaires. Au contraire, cette approche semble redondante en ce qui concerne les modèles non-linéaires et tend à dégrader leur performance en comparaison avec une approche directement basée sur l'intégration de variables exogènes.

Le recours aux variables synoptiques ne permet pas d'améliorer les performances prédictives, que ce soit en considérant uniquement les informations au niveau du parc ou de ses environs. Ceci est probablement dû à la faible profondeur de l'archive considérée qui ne permet pas l'obtention d'analogues possédant un haut degré de similarité.

### Etude préliminaire du conditionnement appliqué aux modèles probabilistes

Une prévision est par essence incertaine. Il est donc pertinent de proposer une quantification de cette incertitude afin de permettre aux décideurs de faire des choix éclairés. Il existe plusieurs façons de générer des prévisions probabilistes. Dans le cadre de ce chapitre

nous faisons le choix de nous tourner vers des modèles de régression probabilistes ayant fait leurs preuves dans la littérature : le modèle AR est remplacé par la régression quantile, alors que le modèle de forêts aléatoires est substitué par son homologue probabiliste ; le modèle des forêts aléatoires quantile.

Nous démontrons que le conditionnement météorologique a un effet positif sur les modèles probabilistes en améliorant la finesse des prévisions, mais tend à dégrader leurs performances comparativement à l'approche basée sur l'injection de variables explicatives. Ce travail préliminaire met en avant que, même si les performances sont légèrement moindres, l'important coût en termes de temps de calcul ne justifie pas le recours au conditionnement dans le cadre des modèles probabilistes.

### Conditionnement par groupes

Enfin le dernier volet de ce chapitre de thèse concerne la comparaison entre la méthode de conditionnement basée sur les analogues et celle basée sur les clusters. Ce dernier point est détaillé dans la Section C.3. En raison de contraintes temporelles, le conditionnement par cluster n'a pu être investigué que de manière superficielle. Quoi qu'il en soit, les premiers résultats montrent que cette approche permet un gain considérable en termes de temps de calcul par rapport à la méthode des analogues qui consiste à ré-entraîner un modèle pour chaque nouvelle simulation tout en atteignant des performances comparables.

# Chapter 7

# Conclusions and Perspectives

> One never notices what has been done; one can only see what remains to be done.
>
> Marie Curie

**Contents**

## 7.1 Motivations

Due to environmental concerns and energy resources depletion, societies are taking action to reduce greenhouse gas emissions. Among these low-carbon strategies, Photovoltaic (PV) generation appears as a promising alternative to carbon-based energies. Since PV generation is weather-dependent, it is characterised by high variability and limited predictability. When PV constitutes a significant share in power systems, these features raise challenges for power system operators, which have to ensure a high level of power quality and strike a balance between production and demand. To deal with the issue of intermittency, Renewable Energy Sources (RES) forecasting appears as a cost-effective option that can anticipate power imbalances and lead to optimal use of flexibility solutions or traditional adjustment means.

Nonetheless, PV forecasting is still considered as immature, as illustrated by the enthusiasm that this topic generates in the academic field. In addition to a clear profusion of incremental improvements symbolised by a high rate of publication (Figure 1.7), some consortiums, such as Smart4RES, aim at achieving a breakthrough by working together to attain an increase of at least 15% in RES forecasting performance [60]. One step towards this goal is to provide very high-resolution RES-dedicated weather forecasting with $10-15\%$ improvement.

This clear and crucial need to improve forecast accuracy constitutes the global objective of this thesis. To achieve this goal, we have noted several research gaps in the literature. Firstly, quality assessment of power measurements mainly consists of a sequence of basic control steps, to the extent that observations are rarely questioned in more detail. Yet, key components of PV farms may suffer from production shutdowns, which artificially reduce the whole production level. In such a case, there is a discrepancy between the production signal and the explanatory features, which may result in reduced accuracy when such data are used to train forecasting models. Secondly, we observe a limitation in forecasting models to exploit large and heterogeneous sources of data. Often, these models are horizon- and data sources-dependent, which may impact their use in an operational context. Yet several studies highlight that a combination of data is key to improving forecast accuracy and to naturally extend the range of horizons. Thirdly, we observe a growing tendency to resort to complex models, which often act like black boxes. In this context, it becomes challenging to understand how the model will behave, which reduces the range of options to improve the model's performances. Fourthly, there is a clear dichotomy between physics- and statistics-based forecasting. The former field uses physical equations to model the way PV parks work, while the latter assumes that statistical models can learn how the plant works on their own, given that relevant information are provided. Our literature review highlighted several research questions, which structure this work:

**RQ1:** How do plants' key components failures impact forecasting accuracy?

**RQ2:** What is the best way to emphasise relevant information contained in irradiance-related fea-

tures?

**RQ3:** What is the optimal methodology to couple several sources of information?

**RQ4:** What strategy can deal with large datasets and horizon-dependent sources of information?

**RQ5:** What is the best way to integrate physics-based information in statistical regression models?

**RQ6:** How can we enhance the interpretability of black-box models?

## 7.2 Summary and main findings

The introductory chapter underlined that the extensive use of carbon-based fuels has had disastrous consequences on wildlife. In this context, RES constitute a desirable option, but their large-scale deployment may jeopardise the safe use of the electrical grid due to their weather dependency. Hence the challenge of developing accurate forecasting tools. However, as pointed out in the previous section, several hurdles must still be overcome.

Chapter 2 details the overarching methodology used throughout this work. First, the different models, which include regression and feature selection models, are presented. Before looking closely at the different ways of improving forecasting performance, it is necessary to know how to quantify a good forecast and develop a relevant verification framework. This verification framework is based on a set of well-established scoring rules and on visual diagnostic tools used or encouraged by the literature. This facilitates forecast analysis and comparison within the literature. In a next step, an overview of the different types of inputs is provided. Preliminary results derived from past production observations are generated. Given the inputs and lead-times considered, these results are in line with what can be observed in the literature.

Usually, PV forecasts rely on statistical models. The latter are employed to infer a large range of processes ranging from the displacement of weather structures to the conversion process occurring within the plant. In chapter 3 we investigate the coupling of physical knowledge with statistical modelling. The main idea is to consider processes that intervene during the conversion of irradiance to electricity (e.g. shading, optical, and even thermal effects) to reduce the computational efforts in the regression model and improve the quality of its fitting. In a nutshell, the different phenomena that may influence the level of power produced (apart from atmospheric conditions) are investigated, and a set of equations involved in the conversion of Global Horizontal Irradiance (GHI) into electrical power is derived. In a next step, the impact of this coupling is analysed from the angle of forecast accuracy. Physics-based conversion is applied either on clear-sky irradiance (i.e. in the context of the clear-sky normalisation) or on irradiance inputs derived from satellite observations or Numerical Weather Predictions (NWPs) models (in this case, clear-sky normalisation is not applied). Results show that in both cases the inclusion of the projection equations improves forecast accuracy, and that the inclusion of equations related to optical effects has a positive impact in the case of normalised inputs, while the influence of thermal effects is more ambiguous.

Once the conversion model has been defined, we can use it to analyse the datasets at our disposal and extract the relevant information they contain. That is the objective of Chapter 4. In a first step, the quality of power measurements is investigated. Key components of PV plants may experience inopportune shutdowns that reduce the level of production in relation to the capacity of the impacted components. This can be viewed as a change of weather and can negatively impact Spatio-temporal (ST) models or the fitting of models based on predicted irradiance. A method based on the use of proxies is implemented to assess the production levels of the park in normal conditions (i.e. without component failures). Results show that removing fallacious observations from the dataset has a positive impact on forecasting performances. The literature highlights that missing data may negatively impact forecast accuracy because they alter the data distribution. That is why correction is considered. However, in our case, the correction step does not improve the accuracy of the forecasts; this may be attributed to noise or artificial correlations generated by the corrective process. In a second step, the clear-sky normalisation of irradiance-related features is investigated. This approach allows us to remove the deterministic trend associated with the Sun's path. We observe that the normalisation of power measurements with power-like features derived from clear-sky irradiance leads to slightly better stationary properties. Machine Learning (ML) techniques are traditionally used with raw features, in the sense that they are not normalised with clear-sky related features. However, the results tend [1] to show a positive impact in terms of forecast accuracy when clear-sky normalisation is used.

The clear-sky normalisation process allows us to remove the misleading ST correlations associated with the dependency of the irradiance on the Sun's path, which makes the integration of spatially distributed sources of information possible. Three types of inputs are investigated: (1) spatially distributed power units, (2) Satellite Derived Surface Irradiance (SDSI), and (3) opacity maps derived from infrared channels. First, in the case of spatially distributed power units, a pre-selection process based on physics-based time decorrelation distance is implemented. This allows us to retain only features from sites that are close enough to experience the same ST structures. Then, a feature selection process, such as the Least Absolute Shrinkage and Selection Operator (LASSO), is implemented to select relevant features. An analysis based on wind directions exhibits that preponderant winds are associated with the north-south axis, which benefits to the spatial distribution of the sites. However, these winds do not carry a lot of clouds, unlike winds from the west. This motivates us to resort to satellite-based observations to deal with the issue of the low density of the plant network and their monotonous distribution along the Rhône river (i.e. north-south axis). To deal with the computational burden induced by this type of information, a new feature selection scheme is implemented. The latter aims at selecting a set of features with the lowest redundancy, while maximising the dependency on the target feature. The forecasts issued with this method outperform the forecasts obtained with a traditional ap-

1. A significant improvement is observed in the case of Random Forest (RF) but no clear distinction is made when an Artificial Neural Networks (ANN) is used.

proach based on the maximisation of a dependency criterion. It is interesting to observe that the structure of the Auto-Regressive (AR) model is not suitable to extract all the relevant information carried by the subset of features derived from SDSI maps, while nonlinear models such as RF are able to do so. With the ambition to improve forecasts generated for the early morning, we consider satellite-based information derived from infrared channels. This source of input is rather under-represented in the literature. This may be explained by the fact that information carried by such data is more difficult to assess than irradiance forecasts. In that regard, the results show that early morning forecasts generated with irradiance forecasts are more accurate. However, the integration of both sources of information (i.e. irradiance forecasts and features derived from opacity maps) leads to higher accuracy.

In a ST context, the importance of features depends on the direction of displacement of cloud structures [62]. Nonetheless, the models considered so far possess static parameters. Therefore, the model fitting mainly reflects the predominant situations encountered during its training step. Hence the idea of dynamically updating these parameters according to the weather states. In this paradigm, models are trained on a batch of data sharing common characteristics with the situation to predict, which leads to adaptive models. This approach is developed throughout Chapter 6, which lays the mathematical foundations of a generic methodology to dynamically update models' parameters. This Weather-Conditioned (WHCO) approach can be viewed as a natural way of including nonlinear capabilities in models. Results show that this approach is well adapted for linear models in the sense that it naturally extends the models' capabilities, especially with features that have nonlinear dependencies on the response variable, while it seems redundant with nonlinear models. Nevertheless, WHCO approaches are used in the literature with a wide range of regression tools. Therefore, special care should be taken regarding the nature of the model. In the case of a linear regression model, the WHCO approach gets the best from features that have nonlinear dependencies on the response variable in comparison with a straightforward integration as explanatory features. However, features with linear dependencies provide better performances when considered as explanatory features. Ultimately, higher scores are reached when both modes are employed simultaneously. The proposed approach is based on the analogy principle and requires fitting a new model for each new forecast. A preliminary work highlighted that shorter computing times are achieved when considering models trained on clusters of data sharing similar weather characteristics. Such an approach provides similar forecast accuracy. Last but not least, a preliminary investigation of the impact of weather conditioning over probabilistic forecasting models is conducted. First results show that WHCO tends to produce sharper forecasts, nonetheless, associated performances are lower than in the case of considering explanatory features. Be that as it may, further works are needed to consolidate these findings.

The different secondary research results provided throughout this document are sum-

marised in Table 7.1.

| **Chapter** (Section) | **Application** | **Learning** |
|---|---|---|
| **Physics-based Modelling** (3.5.1.2) | GHI conversion | Projection of irradiance on the POA improves forecast accuracy. |
| **Physics-based Modelling** (3.5.2) | Irradiance components | Forecasts based on BHI and DHI outperforms forecasts issued from GHI. |
| **Physics-based Modelling** (3.5.2) | Physical knowledge derived from statistical model | A nonlinear model is able to derive conversion laws when fed with relevant inputs. |
| **Data Characterisation** (3.5.2) | Identification of spurious observations | The rejection of fallacious observations leads to an improvement of the model bias and has a positive impact for forecast horizons greater than 2-hour ahead in terms of nRMSE and nMAE. |
| **Data Characterisation** (4.4.2.2.3) | Clear-sky normalisation (ML models) | RF model performs better when fed with clear-sky normalised inputs. |
| **Data Characterisation** (4.4.2.2.3) | Clear-sky normalisation (stationarity) | The normalisation approaches based on the irradiance-projection and performance models seem to exhibit local stationarity properties. |
| **Spatio-temporal Information** (5.3.2.1.3) | Dimensionality issue | The mRMR framework improves forecasting accuracy of both linear and nonlinear models compared to other methods based on the maximisation of a correlation criterion. |
| **Spatio-temporal Information** (5.3.4) | Mixing of data sources | The main source of information is provided by satellite-based information. |
| **Spatio-temporal Information** (5.4.2) | Opacity maps compared with SDSI maps | Opacity maps improve forecasts for the early morning. For daytime-issued forecasts, models fed with this input are outperformed by models based on SDSI. |
| **Spatio-temporal Information** (5.4.2) | Opacity maps compared with NWPs | Irradiance forecasts provide more informative data. However, the combination of both inputs leads to a global accuracy improvement. |

Continued on next page

**Table 7.1 – continued from previous page**

| **Chapter** (Section) | **Application** | **Learning** |
|---|---|---|
| **Conditioned Learning** (6.4.2.2) | Comparison of integration modes | In the case of linear models, the WHCO approach gets the best from features that have nonlinear dependencies on the response variable. Features with linear dependencies perform better when considered as explanatory features, but higher scores are reached when both modes are employed simultaneously. |
| **Conditioned Learning** (6.4.3) | Weather conditioning and model family | The explanatory and state features integration mode is well adapted to linear models. On the contrary, nonlinear models perform better with explanatory features. |
| **Conditioned Learning** (6.5) | Gridded data | The use of geopotential field does not improve forecast accuracy, either in the scope of temporal-based forecasts or with ST models. |
| **Conditioned Learning** (6.6) | Probabilistic forecasts | Weather conditioning tends to impact positively the sharpness of forecasts. Still, models fed with NWPs as explanatory features reach higher performances than models conditioned to the local weather state. |

Table 7.1 – Summary of the main results and associated learning

## 7.3 Main take away messages

Henceforth, we have at our disposal all the information to answer the research questions initially raised by the literature review.

**RQ1: How do plants' key component failures impact forecasting accuracy?**

In this work, we investigated failures at the transformer and inverter levels. Shutdowns of such key components deteriorate the quality of the production signal by introducing a new variability component. After the identification of fallacious observations comes the question of how to deal with them: is it better to remove them or correct them? A literature review highlighted that introducing missing observations can deteriorate the accuracy of forecasts by altering the data distribution. In our case study, we observed that removing fallacious observations from both the learning and testing set improves the quality of forecasts.

**RQ2: What is the best way to emphasise relevant information contained in irradiance-related features?**

We have highlighted that information contained in irradiance features can be emphasised in two ways. First, irradiance-related features can be viewed as a signal composed

of (1) a deterministic component resulting from the Sun's movement in the sky dome, (2) a stochastic one associated with the displacement of weather structures, and, in the case of power production: (3) a stochastic component resulting from abnormal production behaviour. The clear-sky normalisation approach makes it possible to remove the daily and seasonal variability patterns resulting from the Sun's movement, while the identification and imputation strategy cancel shutdown-induced variability. Therefore, the irradiance-related variable is purified, and the relevant information is magnified. Second, we have shown that the clear-sky normalised time series do not fulfil stationary requirements. In other words, variability patterns are still present. Empirical analysis highlighted that the integration of features representing solar angles contributes positively to the assimilation of information carried by irradiance-related features.

**RQ3: What is the optimal methodology to couple several sources of information?**

In this document, we have investigated two ways of integrating information within a regression model. The first one consists in adding the inputs linearly to the model by considering them as explanatory features. The second approach considers data as state features. In this paradigm, a model is fitted on observations for which the state features are similar. Thus, the data are not openly used by the regression model, but still allow the derivation of expert models by implicitly impacting the model's parameters. In this work we compared both approaches under the light of NWPs, but we logically assume that the conclusions drawn can be extended to other sources of information. Results show that the optimal way to couple several sources of information depends highly on the nature of the regression model. In the case of a linear regression model, the WHCO approach gets the best from features that have nonlinear dependencies on the response variable in comparison with a straightforward integration as explanatory features. However, features with linear dependencies perform better when considered as explanatory features. Ultimately, higher scores are reached when both modes are employed simultaneously. In the case of nonlinear models, it is desirable to resort to explanatory features.

**RQ4: What strategy can deal with large datasets and horizon-dependent sources of information?**

In this work we have been confronted with dimensionality issues associated with the use of satellite-derived information. In this specific case, we investigated several approaches to reduce the computational burden. An innovative approach consisted in resorting to the minimal-Redundancy-Maximal-Relevance (mRMR)-based feature selection scheme. This algorithm offers an interesting option to deal with redundant information. In addition, contrary to traditional correlation criteria such as the Pearson correlation, the mRMR approach, being based on the Mutual Information (MI) criterion, has the ability to identify nonlinear relationships. This approach has been used in the context of satellite-based information, but it can be extended to other ST sources of information. In the introductory

chapter, we have seen that the different sources of information are horizon-dependent (Figure 1.8). As such, it is necessary to include a feature selection model to select relevant variables according to the look-ahead time.

**RQ5: What is the best way to integrate physics-based information in statistical regression models?**

In this work, we chose to integrate physics-based knowledge through the pre-processing of inputs. We reviewed the literature dedicated to the conversion of irradiance into electric power to get an idea of the most relevant processes and their associated equations. Efficient conversion laws require good knowledge of the plant's properties, in particular its geometry. Therefore, a conversion model has been produced taking into account available information, computing efforts, and the relative impact of the conversion step. This conversion model intervenes at different stages of the forecasting process. First, it is used to derive electrical power from irradiance observations during the quality assessment of power observations. Second, it can derive estimations of power under clear-sky conditions. Then, these features are used to normalise power production observations, the resulting time series possess better stationary properties than a similar time series obtained with a clear-sky estimation of irradiance. Third, in a context of ML-based models, the conversion of irradiance features into power-like features can achieve higher forecasting accuracy. Results show that the critical conversion steps is the projection of irradiance on the Plane-of-Array (POA). The difference in performances resulting from including temperatures and wind forecasts are of second order. Per se, the normalisation process can be viewed as a way to integrate physics into statistical models inasmuch as equations governing the Sun's movement are used to reduce the complexity of irradiance-related features.

**RQ6: How can we enhance the interpretability of black-box models?**

A model is interpretable when the user understands the way it works. The first step to understand the inner workings of a model is to capture the importance it gives to each feature. As a preliminary step, it is crucial in our opinion to prune irrelevant or redundant features. In the case of satellite-based information, this can be performed with the mRMR feature selection process or the LASSO in a more general way. In the context of black-box models, the WHCO approach is appealing inasmuch as it derives expert models dedicated to a certain type of weather. Even if the inner works of the model are not intelligible, this approach makes it possible to assess the model's behaviour according to the weather situation. To gain knowledge of the driving forces at work in the forecasting model, it is necessary to have a deep understanding of the processes impacting the conversion processes (i.e. shading effect, temperature dependency of the efficiency) and the dynamics of the irradiance variability (i.e. seasonality). Such knowledge gives precious indications on how to improve forecast accuracy, and on the features' dependencies on each other.

## 7.4 Perspectives

This thesis opens up several research directions regarding PV generation forecasting. This section briefly describes some that seem promising.

**Extend the diversity of information sources**

First, works performed during this thesis mainly rely on generic sources of information traditionally used in the literature. In this regard, very short-term forecasts would benefit from the integration of All-Sky Imagers (ASI)-based information. Many works, including this one, have shown the benefits of integrating spatially distributed information. Real-time sharing of data has become possible thanks to the advances in telecommunication and sensor technologies. Therefore, it would be interesting to consider distributed observations in the context of a data sharing market. Confidentiality constraints would require relevant feature selection methods that could be built upon those presented in this document. Similarly, the observations of non-professional weather station networks could be considered. This source of information turns out to be interesting owing to the democratisation of connected personal weather stations, which provides varied and dispersed measurements of physical parameters. In such a case, close attention should be paid to measurement quality.

**Extend the range of forecasts horizons**

The second research direction that comes to mind is to extend the range of forecast horizons investigated. In this work we confined our investigations to short-term horizons ranging from 15-min to 6-hour ahead, but we could extend it to day-ahead forecast. Per se, the use of ST features do not seem relevant for such lead times, but they could be used to improve the degree of similarity between analog situations in the context of weather conditioning models. In that regard, geopotential fields could be reconsidered inasmuch as they are typically used for day-ahead forecasts [68] or higher lead-times [266].

**Resort to sophisticated models**

A limitation of this thesis is that we mainly rely on state-of-the-art regression models. This deliberate choice results mainly from interpretability concerns. Nonetheless, it could be interesting to investigate the forecast accuracy of Deep Learning (DL)-based tools fed with heterogeneous inputs. In such a case, the pre-processing methods developed in this document could be valuable. The ML and DL fields are evolving fast and sophisticated approaches are emerging, including the transformer model [284], and the Generative Adversarial Networks (GAN) [285]. Works performed in [286] suggests that transformer model can be used to extract different levels of correlation between multiple wind farms and exhibit higher accuracy than Long Short-Term Memory (LSTM) networks.

**Derive probabilistic forecasts from analog-based methods**

Given time constraints, the subject of probabilistic forecasts has only been touched on. Nevertheless, the weather conditioning approach provides fertile ground for the generation of probabilistic intervals from bootstrapping-based methods. Such approaches can easily

convert point forecasts into probabilistic forecasts. For instance, we could contemplate using the bootstrap method to derive confidence intervals through the set of past analog PV production observations used to train the regression model. Such an approach assesses the uncertainty associated with the transformation of the weather forecasts into PV production. In addition, it could be interesting to supplement state features with ensemble forecasts to get an insight into the uncertainty of weather forecasts.

## 7.5 Résumé en Français

### Motivations

Afin de lutter contre la crise climatique qui nous touche actuellement, nos sociétés entreprennent diverses actions visant à réduire nos émissions de gaz à effet de serre. C'est dans ce contexte que les énergies renouvelables ont pris leur essor. Contrairement aux énergies carbonées qui sont pilotables, ces dernières sont tributaires des conditions météorologiques et donc par essence intermittentes, ce qui peut avoir des conséquences néfastes sur la stabilité du réseau. Pour anticiper les déséquilibres entre la production et la consommation, la prévision de la production des énergies renouvelables devient alors primordiale. Cependant, contrairement à la prévision de la production éolienne, la prévision de la production PV est encore considérée comme un domaine immature. L'objectif clé de cette thèse réside donc en l'amélioration de la précision de ces prévisions. Pour mener à bien cet objectif, plusieurs questions de recherche ont été définies à partir de l'état de l'art.

### Résumé et principaux résultats

Dans un premier temps, le Chapitre 2 présente la méthodologie utilisée tout au long de ces travaux de recherche. Ainsi, sont présentés les modèles utilisés pour générer les prévisions, les méthodes pour quantifier la précision de ces dernières, ainsi que les données d'entrée considérées. Une analyse préliminaire des prévisions obtenues avec le modèle AR met en avant des performances en accord avec la littérature.

Le domaine de la prévision PV est dominé par les modèles statistiques ou d'apprentissage machine. Dans le Chapitre 3 nous cherchons à étudier le couplage possible entre ces modèles et les connaissances acquises par la physique. L'idée première est de modéliser physiquement le processus de conversion de l'irradiance en puissance électrique et de l'intégrer à la chaîne de prévision afin d'en étudier l'impact sur la précision des sorties. Les résultats montrent que la prise en compte de la projection de l'irradiance sur plan incliné, et dans une moindre mesure des propriétés optiques des différents matériaux, permet d'améliorer les performances prédictives en comparaison avec un modèle basé sur l'irradiance globale sur plan horizontal.

Une fois ce modèle implémenté, il nous est possible d'analyser finement les données à disposition et d'en extraire l'information pertinente. Ceci constitue l'objectif du Chapitre 4. Ainsi, une méthode y est développée afin d'identifier et éventuellement corriger les anoma-

lies de production qui viennent entacher la qualité des données d'entrée. Nous mettons en évidence la plus-value associée à l'identification et la correction des données en termes de performances prédictives. Néanmoins dans notre cas, ces deux processus conduisent à des résultats très proches. Ceci peut s'expliquer par le bruit généré lors de la phase de correction. Dans un second temps, nous portons notre attention sur la normalisation des données par sortie de modèle ciel clair afin de mettre en exergue la composante aléatoire du signal.

Ce processus de normalisation permet de supprimer les corrélations ST associées au mouvement du soleil. Trois types de données ST sont ensuite analysés et comparés au sein du Chapitre 5 : (1) des observations de production PV spatialement distribuées, (2) des estimations de l'irradiance au sol par imagerie satellite, et (3) des observations de l'opacité nuageuse obtenues grâce aux canaux infrarouges des satellites. Une analyse basée sur la direction des vents dominants associés aux déplacements de masses nuageuses significatives met en évidence une inadéquation avec la configuration de notre cas d'étude. C'est pour pallier ce problème que les données d'origine satellite sont considérées. Afin de gérer la forte dimensionalité de cette source de données, plusieurs approches sont comparées : (1) une approche de sélection des variables basée sur un algorithme de minimisation de la redondance et de maximisation de la pertinence de l'information, (2) une réduction de la dimension via une analyse en composante principale, et (3) une estimation via un réseau convolutif de l'irradiance horizontal au niveau du site d'intérêt à partir des précédentes images. La pertinence de chaque méthode de pré-traitement est tributaire du modèle de régression considéré. Enfin, les images obtenues par canaux infrarouges sont considérées dans l'optique d'améliorer les performances prédictives des prévisions générées pendant la nuit, alors qu'aucune observation récente de la situation météorologique n'est disponible. Cette source d'information est sous-représentée dans la littérature, à notre connaissance seulement deux articles scientifiques en traitent [104, 114]. Nous démontrons la pertinence de cette source pour des prévisions générées pendant la nuit et sa complémentarité avec les estimations de l'irradiance au sol par imagerie satellite et les prévisions de l'irradiance.

Dans un contexte d'utilisation de données ST, l'importance des régresseurs est supposée dépendre de la direction de déplacement des masses atmosphériques. D'où l'idée de mettre à jour de manière dynamique les coefficients des modèles de régression selon l'état de l'atmosphère. Deux approches distinctes peuvent être considérées : soit recaler un modèle pour chaque nouvelle prévision, soit proposer un ensemble de modèles dédiés à certaines situations météorologiques. La première approche est retenue et fait l'objet du Chapitre 6. Cette approche de conditionnement selon la situation météorologique peut être vue comme une façon de non-linéariser les modèles. Ainsi, nous démontrons que cette méthodologie de conditionnement est très bien adaptée aux modèles linéaires mais tend à dégrader les performances des modèles non-linéaires. Le principal défaut de cette approche réside dans son temps de calcul très important. La seconde stratégie de conditionnement présentée à l'Annexe C.3 constitue alors une alternative moins chronophage et conduisant, selon les premiers

résultats, à des performances semblables. Enfin, une étude préliminaire du conditionnement appliqué aux modèles probabilistes montre que cette méthodologie tend à produire des prévisions plus fines mais moins précises en comparaison avec des prévisions générées par le même modèle mais en considérant les données NWPs en tant que variables exogènes.

Nous avons désormais tous les éléments pour répondre aux différentes questions de recherche :

**RQ1 : Comment les défauts des composants principaux des centrales PV ont des répercussions sur la précision des prévisions ?**

Les défauts au niveau des onduleurs et des transformateurs détériorent la qualité du signal de production en introduisant une variabilité additionnelle. Dans notre cas, le rejet des observations fallacieuses permet d'améliorer les performances prédictives.

**RQ2 : Quelle est la meilleure approche pour mettre en valeur l'information pertinente contenue dans les séries temporelles de production ou d'irradiance ?**

La normalisation par ciel clair permet de supprimer la variabilité journalière et saisonnière due à la course du soleil et donc de mettre l'accent sur la composante associée à la variation due aux mouvements de masses nuageuses. Néanmoins, ce processus n'est pas suffisant pour rendre les séries temporelles stationnaires. Une étude empirique souligne l'intérêt d'ajouter des variables additionnelles telles que les angles solaires afin de permettre une meilleure assimilation des informations par le modèle.

**RQ3 : Quelle est la stratégie optimale pour coupler plusieurs sources d'information ?**

La meilleure stratégie d'intégration de variables dépend essentiellement du modèle de régression. Dans le cas d'un modèle linéaire, le conditionnement par la situation météorologique permet de tirer le meilleur parti des variables ayant une dépendance non-linéaire avec la variable à expliquer, alors qu'une intégration directe en tant que variable exogène est indiquée dans le cas d'une variable linéaire. Dans le cas d'un modèle non-linéaire, il est préférable de recourir aux variables explicatives.

**RQ4 : Quelle stratégie pour gérer des jeux de données de grande dimension ?**

Dans ces travaux, nous proposons d'utiliser une approche innovante permettant de sélectionner des variables qui minimisent la redondance et maximisent la pertinence de l'information.

**RQ5 : Quelle est la meilleure approche pour intégrer des connaissances physiques dans un modèle statistique ?**

Nous avons fait le choix d'intégrer des connaissances basées sur la physique en prétraitant les données d'entrée. Nous avons donc réalisé une revue bibliographique afin de proposer un modèle de conversion de l'irradiance en puissance électrique qui soit simple et performant. Ce modèle peut être utilisé de trois manières : (1) dans le cadre de l'évaluation de la qualité des données en convertissant l'irradiance observée sur site en puissance électrique, (2) lors de la normalisation en dérivant une puissance électrique théorique ciel clair, et enfin (3) dans

les modèles d'apprentissage machine en convertissant les données d'irradiance en puissance électrique. Dans ces deux derniers cas de figure, l'utilisation du modèle physique conduit à de meilleures performances. L'étape critique de ce processus est la projection sur plan incliné.

**RQ6 : Comment pouvons-nous améliorer l'interprétabilité des modèles de type boîte noire ?** Une étape préliminaire pour comprendre le fonctionnement d'un modèle est de supprimer les variables redondantes et non pertinentes. En ce sens, les différents processus de sélection de variables implémentés tout au long de cette étude répondent à la question soulevée. De plus, la méthodologie de conditionnement à la situation météorologique est séduisante puisqu'elle propose des modèles spécialisés pour chaque type de situation météorologique . Ainsi même si le modèle considéré est une boîte noire, il est possible de comprendre le comportement de ce dernier selon la situation atmosphérique.

## Perspectives

Enfin, cette thèse ouvre différentes directions de recherche. Tout d'abord, il est envisageable d'étendre la diversité des sources d'informations, par exemple, via la prise en compte d'images du ciel au niveau du site ou en considérant des données observées au niveau de stations météorologiques amateures. Il est également concevable d'étendre les horizons de prévisions, e.g. pour le lendemain. Dans ce contexte, la prise en compte des champs géopotentiels pourrait être judicieuse. Dans cette étude, nous nous sommes volontairement restreints à des modèles de prévisions relativement peu complexes. Il serait donc intéressant d'étudier des architectures plus complexes et novatrices telles que les réseaux antagonistes génératifs. Un dernier axe pourrait consister en la génération d'intervalles de confiance à partir de méthodes basées sur les analogues et le bootstraping.

# Appendices

# Appendix A

# Computational Time

This section gathers the computational time (Table A.1) of the main models investigated throughout this thesis. These times do not include the feature selection of SDSI. The computations are performed on a virtual machine composed of 128 Go RAM and 32 central processing units. To reduce the computational time of the forecast generations, forecasts for the different horizons are run in parallel. Models are trained on data from year 2015 and performances are evaluated on year 2016.

The low computational cost of the *CRF(local) + SDSI(t/mRMR)* model compared with the *CAR(local) + SDSI(t-9:t/mRMR)* is due to the fact that the former considers only the last observations associated with the $N_{SDSI} = 10$ selected SDSI features, while the AR-based model also integrates lagged observations (Section 5.2.4.4). The high computational costs of optimised WHCO models result from the fine grid used during the grid-search optimisation (we evaluate 11 potential values of the number of analog situations: $N = \{200, 400, 800, \ldots, 4000\}$).

| **Model** | **Time (hour)** | |
|---|---|---|
| | Without optimisation | With optimisation* |
| *Persistence* | 00:00:10 | |
| *AR* | 00:00:22 | |
| *RF* | 00:00:22 | |
| *AR + SDSI(t-9:t/mRMR)* | 00:01:00 | |
| *RF + SDSI(t/mRMR)* | 00:00:38 | |
| *AR + SDSI(t-9:t/mRMR) + NWPs* | 00:02:45 | |
| *RF + SDSI(t/mRMR) + NWPs* | 00:02:37 | |
| *CAR(local)* | 00:07:55 | 00:29:01 |
| *CRF(local)* | 00:06:02 | 01:00:37 |
| *CAR(local) + SDSI(t-9:t/mRMR)* | 00:19:49 | 01:01:58 |
| *CRF(local) + SDSI(t/mRMR)* | 00:06:11 | 01:16:37 |
| *CAR(local) + SDSI(t-9:t/mRMR) + NWPs* | 00:23:15 | 01:15:45 |
| *CRF(local) + SDSI(t/mRMR) + NWPs* | 00:08:35 | 01:40:04 |
| *CAR(SSRD)* | 00:07:51 | 00:27:08 |
| *cCAR(SSRD)* | 00:00:30 | |

Table A.1 – Summary of typical computational times (training and testing) of the main forecasting architectures used in this thesis. Time should be understood as the time needed to forecast the production of a site for the various lead times considered (namely $h \in \{15, 30, 45, 60, 75, 90, 105, 120, 180, 240, 300, 360\}$). *The optimisation of the number of analog situations is developed in Section 6.3.1.3.

# Appendix B

# Supplementary material for clear-sky normalisation

## B.1 Choice of the clear-sky model

During this thesis, two clear-sky models have been investigated. To determine the best model, we run the AR forecasting model with PV production normalised by the two models, namely the European Solar Radiation Atlas (ESRA) [207] and the McClear [212] models. These two models mainly differ in the way that they consider aerosols: the ESRA model is based on a monthly climatology, while the McClear model is supplied with parameters updated every 3 hours. As this work was performed in the early stage of this thesis, the influence of the power conversion (i.e. the model that converts GHI into electrical power) was not investigated.

### B.1.1 Forecasting performances

Figure B.1 represents the forecasting performances of the AR model fed with production observations normalised through either the McClear or the ESRA model outputs. In addition, the impact of the projection of GHI on the POA is also assessed. First, it is obvious that whatever the clear-sky model considered, the normalisation approach has a positive impact on the forecasting accuracy. We observe that the forecasting performances obtained with inputs normalised by the McClear model are better in terms of normalised Root Mean Square Error (nRMSE) (i.e. the *AR(McClear-GHI)* model outperforms the *AR(ESRA-GHI)* model), but conclusions are less straightforward when considering the normalised Mean Absolute Error (nMAE). Indeed, for forecast horizons lower than 2-hour ahead, the *AR(ESRA-GHI)* model is better than the *AR(McClear-GHI)* model. This observation is partly in line with the statement made in [107], which stipulates that the Linke Turbidity (LT) factor – which is used in the ESRA model – is a source of uncertainty. In addition, the projection of the GHI improves the nRMSE scores for both clear-sky models, while it tends to degrade the nMAE scores.

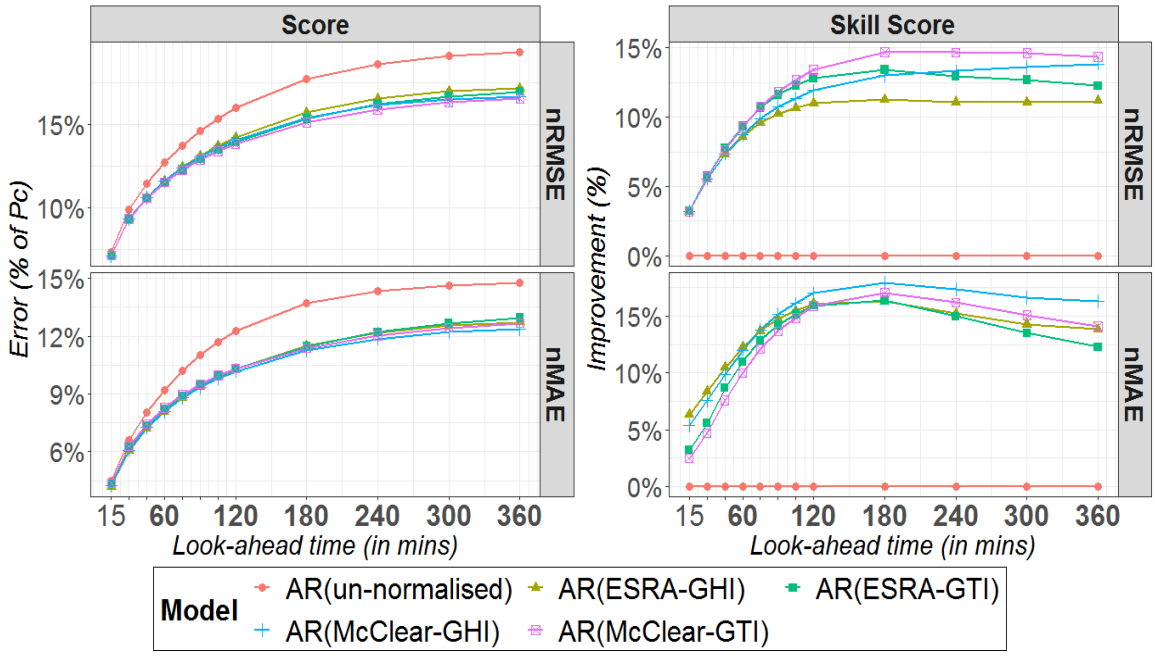



Figure B.1 – Forecasting performances obtained with the AR model fed with production normalised either with the McClear or the ESRA clear-sky models outputs. Influence of the projection is also investigated. Various configurations are assessed: (1) the *AR(un-normalised)* model fitted on non-normalised inputs, (2) the *AR(ESRA-GHI)*, and (3) the *AR(ESRA-GTI)* models are fed with inputs normalised with ESRA-based outputs, and lastly (4) the *AR(McClear-GHI)*, and (5) the *AR(McClear-GTI)* models are fed with inputs normalised by the McClear model outputs.

All the regression models considered in this thesis aim at minimising the Root Mean Square Error (RMSE) between observations and forecasts. Therefore, production observations are then normalised through clear-sky estimations provided by the McClear model.

## B.2 Influence of clear-sky normalisation over an artificial neural network

In Section 4.4.2.2.3, we assess the influence of the clear-sky normalisation process over the accuracy of RF-based models. As a complement, here we focus on an ANN architecture [287] implemented with the keras package [288]. This architecture is composed of 5 layers with respectively 256, 128, 64, 32, and 1 neurons. The Rectified Linear Unit (ReLU) function is chosen as the activation function in all of the layers, except the last one, where a linear function is used. To avoid overfitting, the L2 regularisation penalty is added to the loss function, namely the mean squared error. The Adam optimiser is used.

Similarly to what have been done with the RF model, we compare the forecasting performances of the ANN model fed with clear-sky normalised or non-normalised inputs. To help the model to better assess the influence of the Sun's movement on the signal variability, we add features such as solar angles (elevation and azimuth angles). All the accuracy scores are gathered in Figure B.2. First, we observe that the poorest accuracy is obtained with the

model fed with non-normalised power measurements. The addition of solar angles features improves performances both in terms of nMAE and nRMSE. Nonetheless, the *ANN(k3) + Solar angles* model slightly outperforms the *ANN(PV) + Solar angles* model. In addition, when we consider Surface Solar Radiation Downwards (SSRD) predictions, we observe that the model based on the clear-sky normalisation approach (i.e. the *ANN(k3) + Solar angles + k(SSRD)* model), and the model fed with non-normalised inputs (i.e. the *ANN(PV) + Solar angles + SSRD* model) lead to rather similar accuracy.

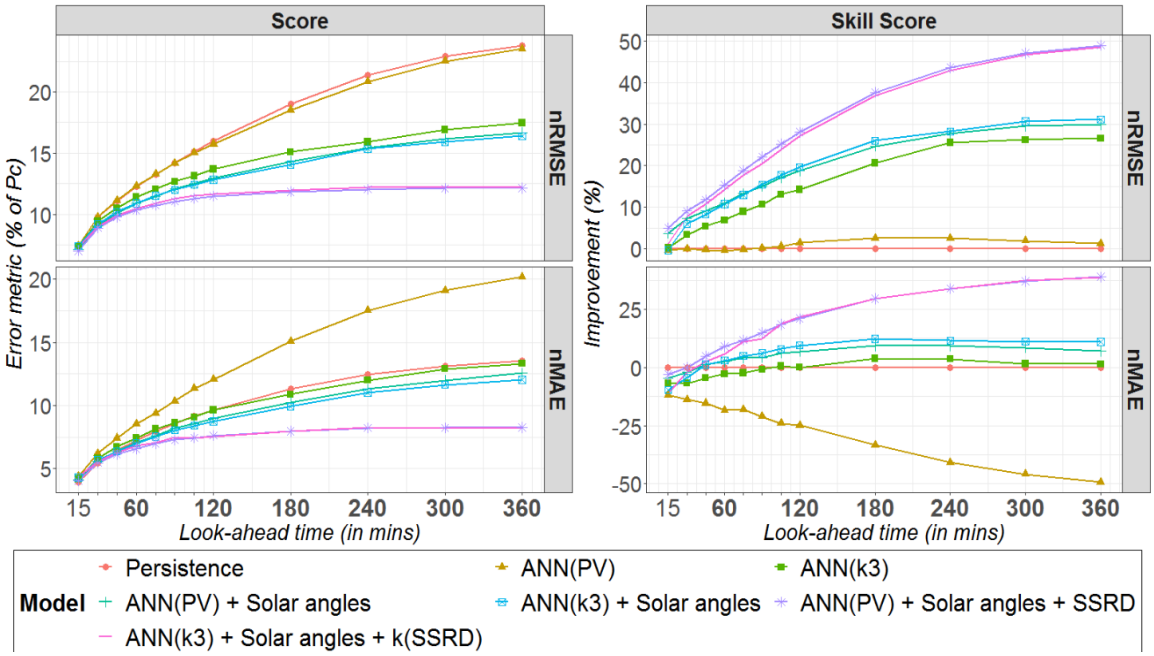



Figure B.2 – Forecasting performances of the ANN model fed with clear-sky normalised inputs (i.e. *ANN(k3)*, *ANN(k3) + Solar angles*, *ANN(k3) + Solar angles + k(SSRD)*) or non-normalised inputs (i.e. *ANN(PV)*, *ANN(PV) + Solar angles*, *ANN(PV) + Solar angles + SSRD*).

As the nRMSE and nMAE differences are very low between the considered models, we implement the Diebold-Mariano (DM) test to judge the statistical significance of the differences. Figure B.3 highlights that for very short-term lead times (typically horizons lower than 1-hour ahead), forecasts produced by the *ANN(k3) + Solar angles + k(SSRD)* and *ANN(PV) + Solar angles + SSRD* models are statistically different. On the contrary, for horizons greater than 3-hour ahead, the models are not statistically different.

To conclude, similarly to the RF model, the use of clear-sky normalised inputs is relevant when past power observations are considered. The use of additional information in relation to the Sun's path helps the model based on non-normalised inputs. The main distinction between the RF and ANN models lies in the fact that the latter is able to make a better use of irradiance predictions in a context of non-normalised data, in such a way that it reaches similar performances than its counterpart based on clear-sky normalised inputs.

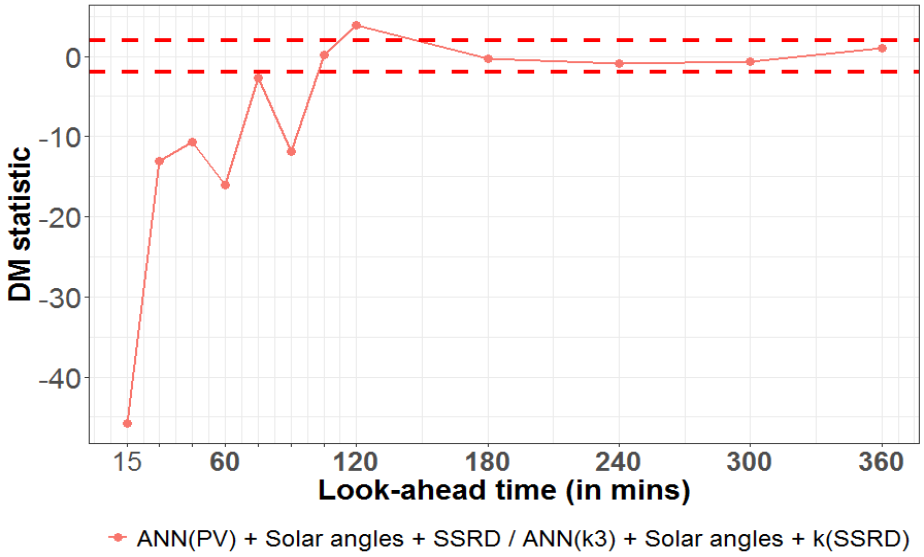


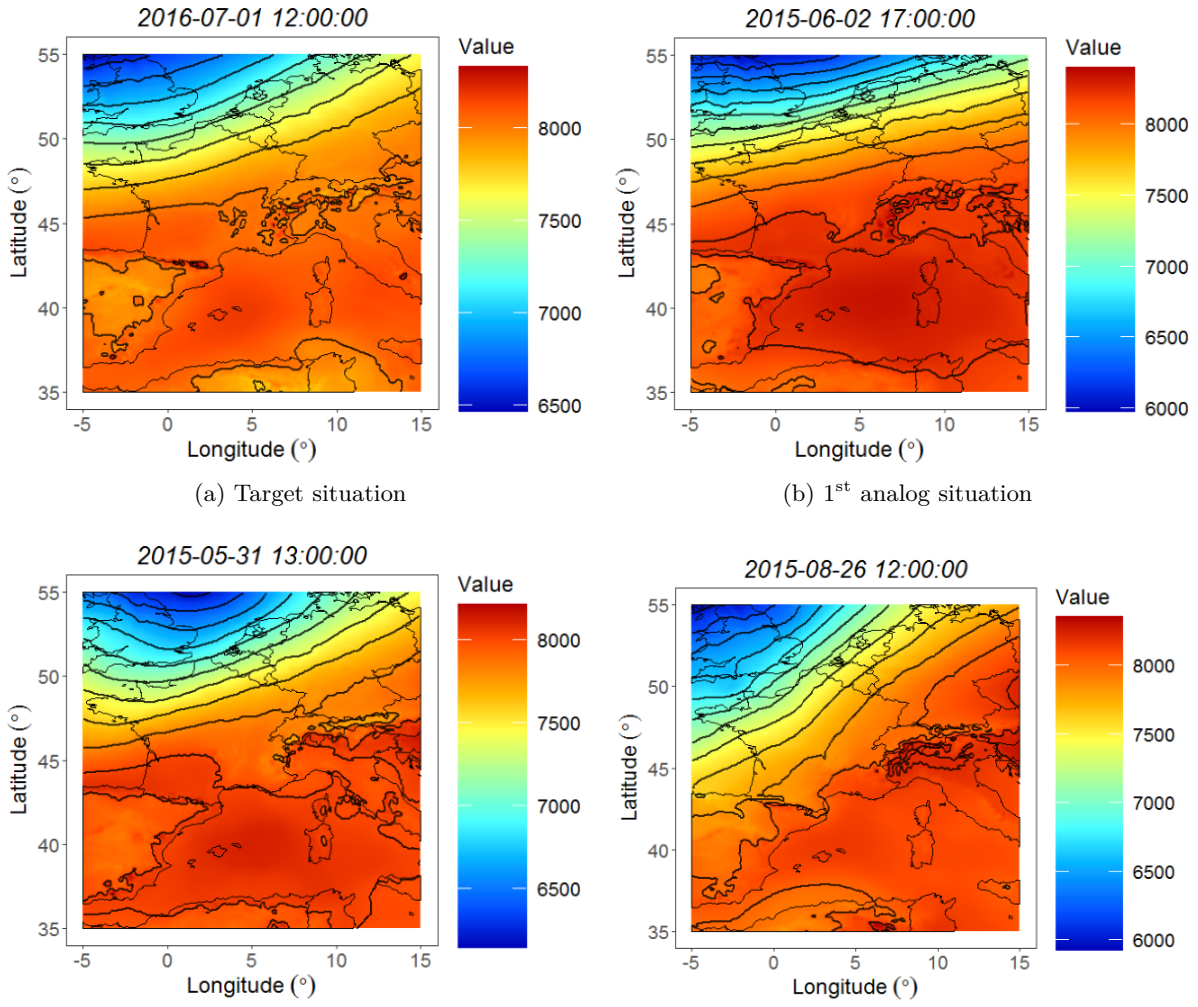

Figure B.3 – DM test (defined in Section 2.3.3) between the *ANN(k3) + Solar angles + k(SSRD)* and *ANN(PV) + Solar angles + SSRD* forecasting models for different forecast horizons. The red dotted lines stand for the borders delimiting the validation and rejection of the null hypothesis.

# Appendix C

# Supplementary material for conditioned learning

## C.1 Analogs obtained with the S1 score

Figure C.1 represents analog situations obtained by considering the score $S_1$ defined in Section 6.3.3.2.2. These sets of figures are to be compared with analog situations (Figure 6.4) obtained with the coupling of the Principal Component Analysis (PCA) approach (Section 5.3.2.2) and the analogy score $D$ (Equation 6.2).




(a) Target situation

(b) 1st analog situation

(c) 10th analog situation

(d) 100th analog situation

Figure C.1 – Examples of analog situations (**b**), (**c**), (**d**) with regard to the target situation (**a**) obtained with the 925 hPa geopotential field with the $S_1$ score.

## C.2 Sensitivity analysis of geopotential fields

This section aims at determining the geopotential field level and the associated spatial window which best characterise the atmospheric state at PV units location. To do so, we adopt an a posteriori analysis framework: forecasting performances of the root model conditioned to various state features are compared, and only the ones leading to the best performances are retained. Figures C.2 and C.3 show the performances of the AR model conditioned respectively with geopotential fields at 500 and 925 hPa for three spatial windows. For the 500 hPa pressure level, we observe that the nRMSE- and nMAE-based skill scores are similar for forecast horizons lower than 2-hour ahead, independently of the spatial window. Beyond, forecasts based on the spatial window *W1* stand out in terms of nRMSE and nMAE scores. As regards forecasts issued with the geopotential field at the 925 hPa pressure level, the best scores are obtained considering the window *W3*.

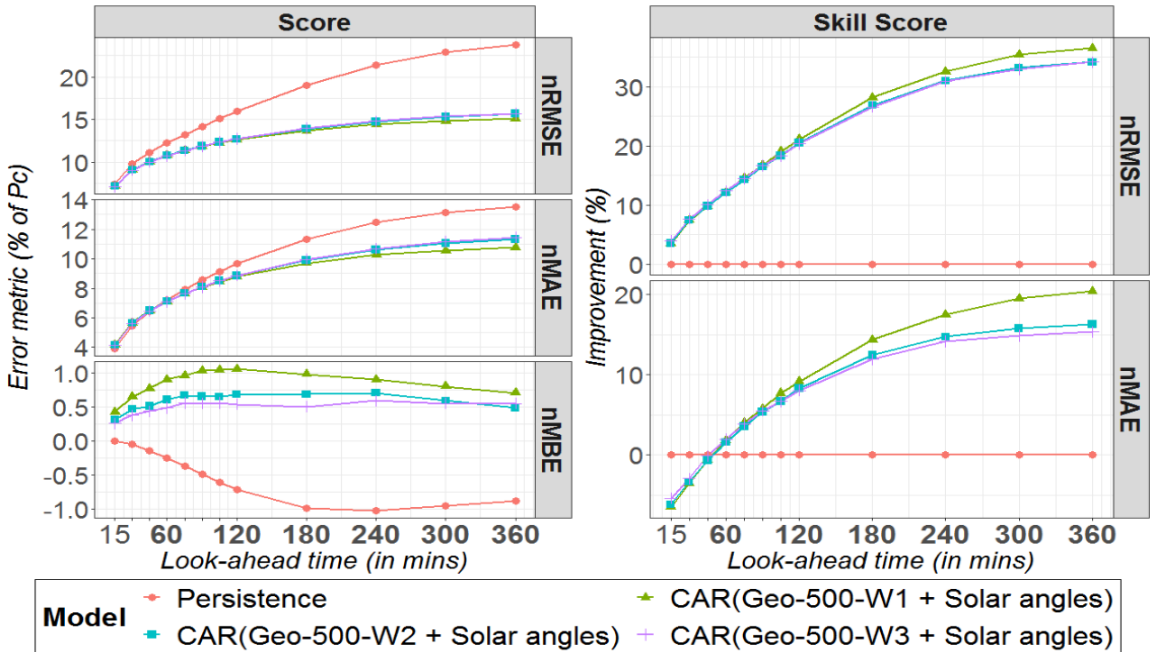



Figure C.2 – Forecasting performances of the AR model conditioned to the geopotential field at 500 hPa and the solar angles for the three spatial windows defined in Figure 6.5.

A comparison between the most accurate forecasts obtained with the 500 and 925 hPa pressure levels is displayed in Figure C.4. Both approaches provide similar scores, but forecasts based on the 925 hPa pressure level are slightly better. The DM test reveals that generated forecasts are statistically different (Figure C.5).

## C.3 Weather conditioning based on clusters

WHCO, as it is defined in Section 6.2.1, can take at least two forms; either a regime-switching model approach, where each model is dedicated to a specific weather type (e.g.

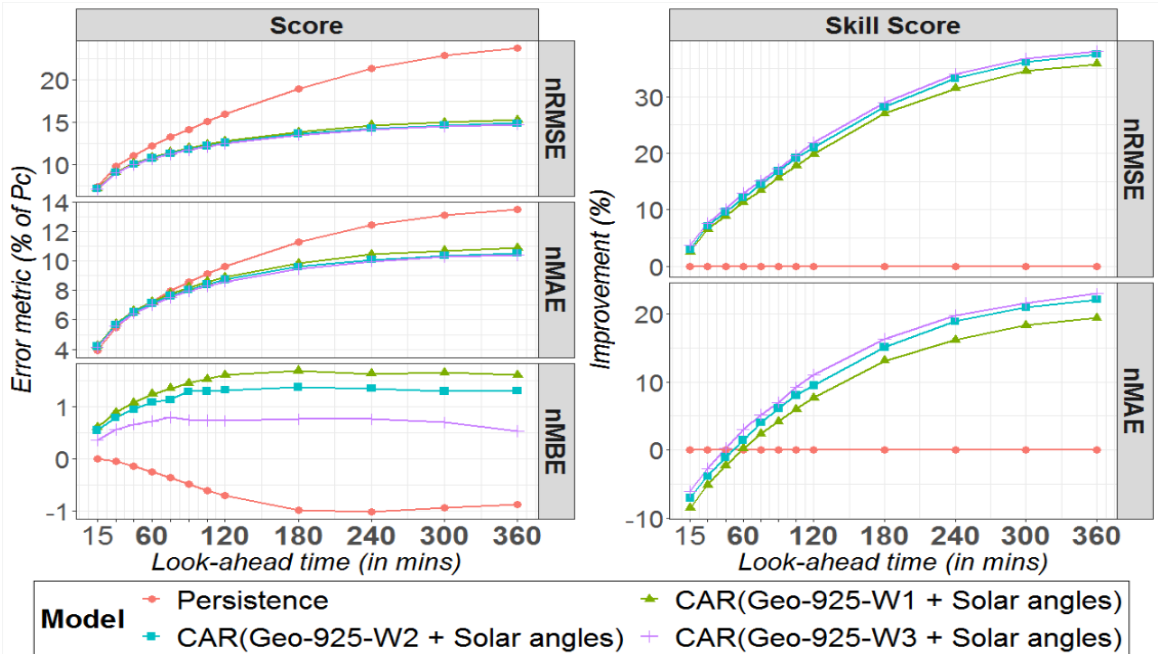



Figure C.3 – Forecasting performances of the AR model conditioned to the geopotential field at 925 hPa and the solar angles for the three spatial windows defined in Figure 6.5.

sunny or cloudy), or a dynamic approach, where the model parameters are updated regularly. In this section, we adopt the former approach. Due to time constraints, the subject is only touched upon, but the proposed methodology can be viewed as a guide for further investigations.

### C.3.1 Methodology

The definition of the various atmospheric regimes is performed by a clustering algorithm. The goal of such an algorithm is to divide $N$ points from a $D$-dimensional space into $K$ clusters in such a way that observations belonging to a same group share more similarity to each other to those in other clusters.

The clustering space in which the determination of clusters is performed is inspired from the analogy score defined in Equation C.1, formerly introduced in Section 6.3.1.2. This metric possesses the property to attribute low scores for weather situations that have parameters with close values and that evolve similarly. To keep these characteristics, we define the clustering space as a combination of features $z_{i,t}$ and their corresponding lagged and leading values: $\mathcal{C} = \{z_{i,t-\widetilde{t}}, z_i, z_{i,t+\widetilde{t}}\}$. In other words, we build a space composed of a set of weather forecast parameters at time $t-\widetilde{t}$, $t$, and $t+\widetilde{t}$ (where $\widetilde{t} = 60$ min). As a result, situations associated with weather parameters that have similar values and evolve likely should be located in the same area of the clustering space. As a preliminary work, the focus is on the SSRD feature normalised through the output of the clear-sky model.

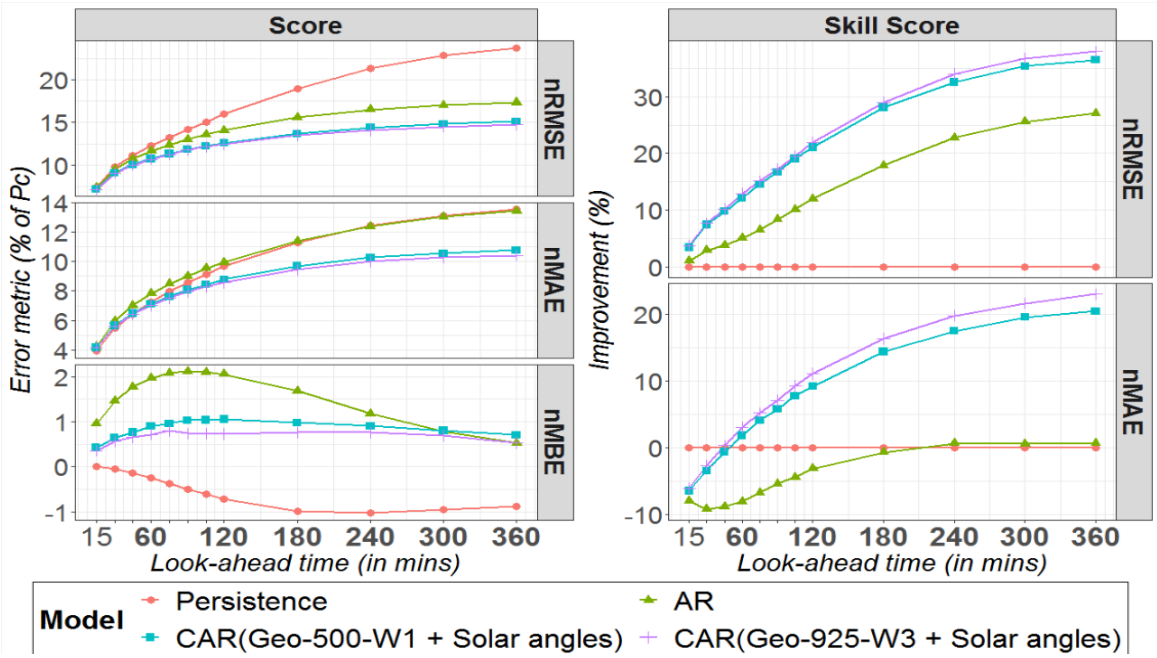



Figure C.4 – Comparison of forecasts obtained with AR model conditioned to geopotential fields at 500 and 925 hPa pressure levels and associated respectively with the spatial windows *W1* and *W3*.

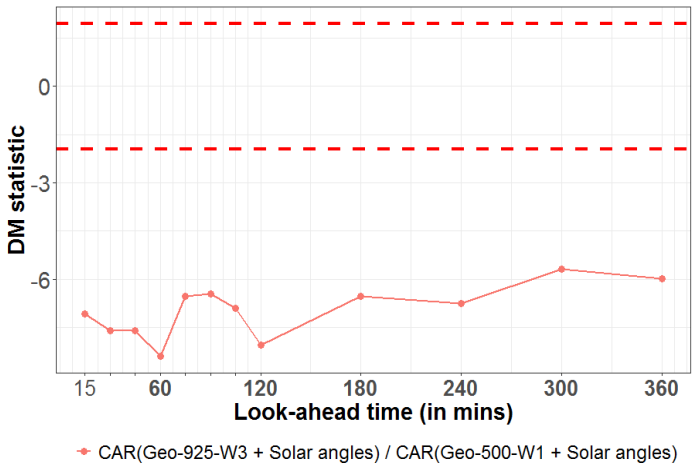



Figure C.5 – DM test (defined in Section 2.3.3) between forecasts obtained with AR model conditioned to the geopotential fields at 500 and 925 hPa. The red dotted lines stand for the borders delimiting the validation and rejection of the null hypothesis.

$$D(Z_{t+h}, Z_{t'+h}) = \sum_{i=1}^{D} \frac{\omega_i^A}{\sigma_i} \sqrt{\sum_{j=-\widetilde{t}}^{\widetilde{t}} (z_{i,t+h+j} - z_{i,t'+h+j})^2}. \tag{C.1}$$

| | |
|---|---|
| $t$ | Moment when the forecast is generated, |
| $t'$ | Temporal observations from the learning set, |
| $i$ | Index referring to the analog predictors, |
| $D$ | Number of analog predictors, |
| $\omega_i^A$ | Weight of analog predictors ($\sum_{i=1}^{N_v} \omega_i^A = 1$), |

$\sigma_i$ Standard deviation of analog predictors,

$\widetilde{t}$ Half-width of the time window over which the metric is computed ($\widetilde{t} = 60$ min).

Several definitions of cluster are proposed in the literature, which leads to the development of specific algorithms (e.g. distance- or density-based algorithms). Here, two algorithms are investigated, namely the Partitioning Around Medoids (PAM) [289] and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [202] algorithms. In short, the PAM (or K-medoids) algorithm is a clustering algorithm very similar to the K-means algorithm [290]. K-medoids [1] is a partitional algorithm that minimises the distance (here, the Euclidean distance) between points from a cluster and the point designated as the centre of that cluster (a medoid in the case of the K-medoids algorithm or a centroid in the case of the K-means algorithm). The PAM algorithm is a more robust version of the K-means regarding noise and outliers because the latter tends to move the centre of the cluster towards the outliers, which in turn move other points away from the centre of the cluster. On the other hand, DBSCAN algorithm relies on the notion of density: within each cluster, the density of points is greater than outside the cluster and the density within an area of noise is lower than the density in any of the clusters. The key idea behind DBSCAN is that for each point of a cluster the neighbourhood of a given radius, $\epsilon$, has to contain at least a minimum number of points denoted as $MinPts$. This algorithm is detailed in Section 4.3.3. Both algorithms need user-defined parameters (e.g. the number of clusters $K$ or a minimal neighbourhood radius $\epsilon$). These parameters are determined through a grid search approach where several parameters are tested and only those leading to the best forecasting performances are retained. A comparison between forecasts issued by the best-tuned versions of the PAM and the DBSCAN algorithms highlights that the former provides the best accuracy when the AR model is used. This is supposed to result from the lack of clear demarcation between the different regimes observed in the clustering space. Indeed, Figure C.6 highlights that a density-based clustering approach is not fitted to the data we are dealing with: no clear density variation is observed within the distribution of points along the diagonal. This figure shows regimes with specific weather dynamics. For instance, cluster 5 stands for sunny situations, cluster 14 characterises overcast situations, while cluster 6 corresponds to crepuscular situations.

1. A medoid is a representative object of a dataset, in contrast to the centroid, or centre of mass that may not necessarily belong to the dataset.

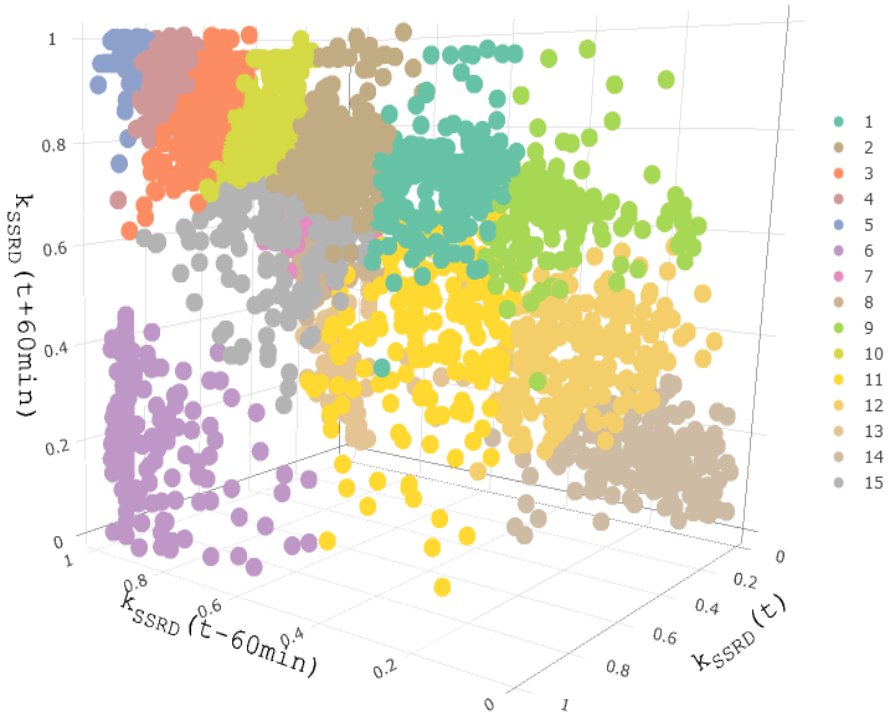



Figure C.6 – 3-D scatter plot of clusters obtained with the PAM algorithm considering a space composed of the SSRD normalised by clear-sky model outputs. Colours represent the 15 clusters.

One of the weaknesses of PAM algorithm is that it processes all features equally, independently of their actual relevance. This issue can be circumvented by considering a weighted distance as a dissimilarity metric (such as the weighted Euclidean distance, Equation C.2). Then a grid search approach performed on the weights can be implemented to obtain the tuning leading to the best forecasting accuracy. In order to simplify the process of determining relevant weight for each features while reducing the computational time, we can derive weights from the MI criterion (notion defined in Section 5.3.2.1.1) between the features and the PV production observations as in [67]. Nonetheless, due to time constraints none of these solutions have been investigated.

$$d(z_i, m_k) = \sum_{d=1}^{D} \omega_d (z_{i,d} - m_{k,d})^2 \tag{C.2}$$

| | |
|---|---|
| $D$ | Number of features considered in the clustering analysis, |
| $k$ | Considered cluster, |
| $m$ | Medoid of cluster k, |
| $\omega_d$ | Feature specific weight, |
| $d(z_i, m_k)$ | Dissimilarity measure between entity $z_i$ and medoid $m_k$. |

The determination of the clusters is performed on the training dataset, then characteristics of the determined clusters are used to assign entities in the testing dataset. Eventually,

$D$ regression models are fitted on the clusters of the training set, and then applied to the corresponding clusters of the testing dataset.

### C.3.2 Results

The grid search regarding the number of clusters $K$ to consider is performed with the AR model. The best forecasting accuracy is reached when considering 15 clusters, beyond, improvements are negligible. Figure C.7 gathers the forecasting performances of analog-[2] and cluster-based conditioned models. When SSRD is used only as a state feature, the cluster-based conditioned (cC) model (i.e. the *cCAR(SSRD)* model) exhibits the best scores compared to the analog-based model (i.e. *CAR(SSRD)*) in terms of nRMSE, nMAE, and normalised Mean Bias Error (nMBE). Apart from providing forecasts with higher accuracy, the cluster-based conditioning drastically reduces the computational time: it is around 50 times faster than the analog-based conditioning (Table A.1). The simultaneous use of SSRD as a state and explanatory feature leads to ambivalent scores. Both models exhibit rather similar nMAE scores, but the *CAR(SSRD)+SSRD* model slightly outperforms its counterpart in terms of nRMSE, while the *cCAR(SSRD)+SSRD* model has slightly a lower bias. Differences among these forecasts are significant for forecast horizons greater than 3-hour ahead (Figure C.8).

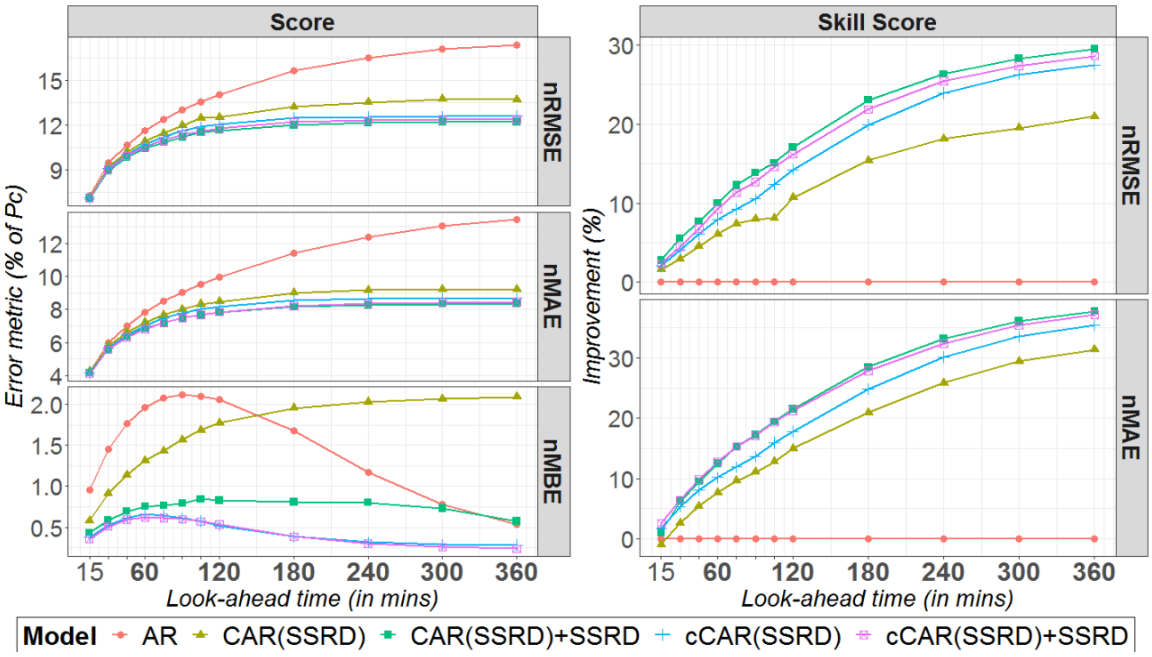



Figure C.7 – Comparison between forecasts issued by analog-based conditioning (i.e. *CAR(SSRD)* and *CAR(SSRD)+SSRD*) and cluster-based conditioning (cC) (i.e. *cCAR(SSRD)* and *cCAR(SSRD)+SSRD*) models.

2. Analog-based conditioning refers to the methodology investigated in Chapter 6 and defined in Section 6.3.

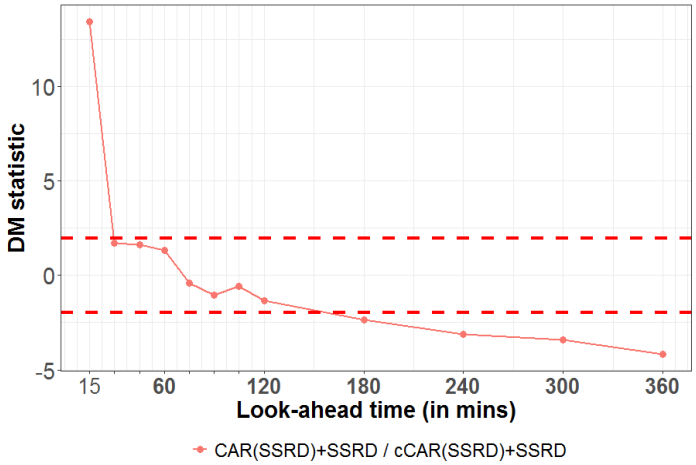



Figure C.8 – DM test (defined in Section 2.3.3) between forecasts obtained with the AR model conditioned to the SSRD either with the analog-based approach or the cluster-based approach. The red dotted lines show the borders delimiting the validation and rejection of the null hypothesis.

### C.3.3 Conclusions

The cluster-based conditioning approach appears as an appealing alternative to the analog-based conditioning. Both approaches exhibit rather close forecasting performances, but the main advantage of the cluster-based conditioning lies in its frugality in terms of computing resources. Further investigations are needed regarding the integration and the weighting of additional state features.

# Bibliography


[1] Nima Norouzi, Maryam Fani, and Zahra Karami Ziarani. The fall of oil Age:A scenario planning approach over the last peak oil of human history by 2040. *Journal of Petroleum Science and Engineering*, 188:106827, May 2020. ISSN 0920-4105. , 10.1016/j.petrol.2019.106827. URL https://www.sciencedirect.com/science/article/pii/S092041051931246X.

[2] Robert Hirsch, Roger Bezdek, Robert Misi, Wendling, and Misi. Peaking of World Oil Production: Impacts, Mitigation, and Risk Management. *Technical Report*, January 2005. , 10.2172/939271.

[3] V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, Y. Chen, L. Goldfarb, M. Gomis, J.B.R. Matthews, S. Berger, M. Huang, O. Yelekçi, R. Yu, and B. Zhou. Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, IPCC, 2021. URL https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM_final.pdf.

[4] Cheryl A. Logan, John P. Dunne, C. Mark Eakin, and Simon D. Donner. Incorporating adaptive responses into future projections of coral bleaching. *Global Change Biology*, 20(1): 125–139, January 2014. ISSN 13541013. , 10.1111/gcb.12390. URL https://onlinelibrary.wiley.com/doi/10.1111/gcb.12390.

[5] A. J. McMichael and World Health Organization, editors. *Climate change and human health: risks and responses*. World Health Organization, Geneva, 2003. ISBN 978-92-4-156248-5.

[6] Lena Fischer, Nejla Gültekin, Marisa B. Kaelin, Jan Fehr, and Patricia Schlagenhauf. Rising temperature and its impact on receptivity to malaria transmission in Europe: A systematic review. *Travel Medicine and Infectious Disease*, 36:101815, July 2020. ISSN 1477-8939. , 10.1016/j.tmaid.2020.101815. URL https://www.sciencedirect.com/science/article/pii/S1477893920303112.

[7] European Environment Agency. *Air quality in Europe: 2020 report.* Publications Office, LU, 2020. URL https://data.europa.eu/doi/10.2800/786656.

[8] O Edenhofer, R Pichs-Madruga, and Y Sokona. Climate Change 2014: Mitigation of Climate Change.Contribution of Work-ing Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, IPCC, United Kingdom and New York, NY, USA, 2014. URL https://www.ipcc.ch/site/assets/uploads/2018/02/ipcc_wg3_ar5_summary-for-policymakers.pdf.

[9] Legifrance. LOI n° 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte (1), 2015. URL https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031044385/.

[10] Legifrance. LOI n° 2019-1147 du 8 novembre 2019 relative à l'énergie et au climat (1), November 2019.

[11] Ministry for the ecological and solidary transition. National Low Carbon Stratetegy. Technical report, French government, March 2020. URL https://www.ecologie.gouv.fr/sites/default/files/en_SNBC-2_complete.pdf.

[12] Ministry for the ecological and solidary transition. French Strategy for Energy and Climate - Multi Annual Energy Plan, 2019. URL https://www.ecologie.gouv.fr/sites/default/files/15.%20PPE%20-English%20Full%20document%20for%20public%20consultation.pdf.

[13] RTE. RTE Bilan Electrique 2020, 2021. URL https://assets.rte-france.com/prod/public/2021-03/Bilan%20electrique%202020_0.pdf.

[14] IRENA. IRENA - Query Tool, April 2020. URL https://www.irena.org/Statistics/Download-Data.

[15] Zengxun Liu, Yan Zhang, Ying Wang, Nan Wei, and Chenghong Gu. Development of the interconnected power grid in Europe and suggestions for the energy internet in China. *Global Energy Interconnection*, 3(2):111–119, April 2020. ISSN 2096-5117. , 10.1016/j.gloei.2020.05.003. URL https://www.sciencedirect.com/science/article/pii/S2096511720300451.

[16] ENTSO-E. The 50 Year Success Story – Evolution of a European Interconnected Grid, 2003. URL https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/ce/110422_UCPTE-UCTE_The50yearSuccessStory.pdf.

[17] M. Safiuddin. History of Electric Grid. In *Foundations of Smart Grid*, pages 6–11. Pacific Crest, January 2013. ISBN 978-1-60263-070-3.

[18] World Nuclear. Nuclear Power in France | French Nuclear Energy - World Nuclear Association, January 2021. URL https://www.world-nuclear.org/information-library/country-profiles/countries-a-f/france.aspx.

[19] ENTSOE. ENTSO-E at a glance. URL https://www.entsoe.eu/publications/general-publications/at-a-glance/.

[20] EURELECTRIC and ENTSO-E. Deterministic frequency deviations –root causes and proposals for potential solutions, December 2011. URL https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/entsoe/120222_Deterministic_Frequency_Deviations_joint_ENTSOE_Eurelectric_Report__Final_.pdf.

[21] Commission de régulation de l'énergie. Services système et mécanisme d'ajustement, November 2019. URL https://www.cre.fr/Electricite/Reseaux-d-electricite/services-systeme-et-mecanisme-d-ajustement.

[22] RTE. Respond to the manual frequency restoration reserve and replacement reserve calls for tenders - RTE Services Portal. URL https://www.services-rte.com/en/learn-more-about-our-services/respond-to-the-manual-frequency.html.

[23] Schéma décennal de développement du réseau - Chapitre 03. Technical report, RTE, 2019. URL https://assets.rte-france.com/prod/public/2020-07/SDDR%202019%20Chapitre%2003%20-%20Les%20adaptations.pdf.

[24] Transport & Environment. CO2 Emissions From Cars: The Facts. Technical report, April 2018. URL https://www.transportenvironment.org/sites/te/files/publications/2018_04_CO2_emissions_cars_The_facts_report_final_0_0.pdf.

[25] IEA. Global EV Outlook 2021. Technical report, IEA, Paris, 2021.

[26] Hasan Mehrjerdi and Elyas Rakhshani. Vehicle-to-grid technology for cost reduction and uncertainty management integrated with solar power. *Journal of Cleaner Production*, 229: 463–469, August 2019. ISSN 0959-6526. , 10.1016/j.jclepro.2019.05.023. URL https://www.sciencedirect.com/science/article/pii/S0959652619315392.

[27] Nuh Erdogan, Fatih Erden, and Mithat Kisacikoglu. A fast and efficient coordinated vehicle-to-grid discharging control scheme for peak shaving in power distribution system. *Journal of Modern Power Systems and Clean Energy*, 6(3):555–566, May 2018. ISSN 2196-5420. , 10.1007/s40565-017-0375-z.

[28] Gheorghe Badea, George Sebastian Naghiu, Ioan Giurca, Ioan Aşchilean, and Emanuel Megyesi. Hydrogen Production Using Solar Energy - Technical Analysis. *Energy Procedia*, 112:418–425, March 2017. ISSN 1876-6102. , 10.1016/j.egypro.2017.03.1097. URL https://www.sciencedirect.com/science/article/pii/S1876610217312225.

[29] B. Lyseng, T. Niet, J. English, V. Keller, K. Palmer-Wilson, B. Robertson, A. Rowe, and P. Wild. System-level power-to-gas energy storage for high penetrations of variable renewables. *International Journal of Hydrogen Energy*, 43(4):1966–1979, January 2018. ISSN 0360-3199. , 10.1016/j.ijhydene.2017.11.162. URL https://www.sciencedirect.com/science/article/pii/S0360319917345809.

[30] Feras Alshehri, Víctor García Suárez, José L. Rueda Torres, Arcadio Perilla, and M. A. M. M. van der Meijden. Modelling and evaluation of PEM hydrogen technologies for frequency ancillary services in future multi-energy sustainable power systems. *Heliyon*, 5 (4):e01396, April 2019. ISSN 2405-8440. , 10.1016/j.heliyon.2019.e01396. URL https://www.sciencedirect.com/science/article/pii/S2405844018367471.

[31] Thomas Carriere, Christophe Vernay, Sebastien Pitaval, Francois-Pascal Neirac, and George Kariniotakis. Strategies for combined operation of PV/storage systems integrated into electricity markets. *IET Renewable Power Generation*, 14(1):71–79, 2020. ISSN 1752-1424. , https://doi.org/10.1049/iet-rpg.2019.0375. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2019.0375.

[32] Scott Burger, Jose Pablo Chaves-Ávila, Carlos Batlle, and Ignacio J Pérez-Arriaga. The Value of Aggregators in Electricity Systems, 2016.

[33] magnuscmd. LIQUIDITY and its Impact on ENERGY MARKETS, May 2016. URL https://www.magnuscmd.com/liquidity-and-its-impact-on-energy-markets/.

[34] Federal Ministry for Economic Cooperation and Development. Variable Renewable Energy Forecasting – Integration into Electricity Grids and Markets – A Best Practice Guide. Technical report, Federal Ministry for Economic Cooperation and Development, March 2015. URL https://energypedia.info/images/2/2a/Discussion_Series_06_Technology_web.pdf.

[35] J. Antonanzas, D. Pozo-Vázquez, L. A. Fernandez-Jimenez, and F. J. Martinez-de Pison. The value of day-ahead forecasting for photovoltaics in the Spanish electricity market. *Solar Energy*, 158:140–146, December 2017. ISSN 0038-092X. , 10.1016/j.solener.2017.09.043. URL http://www.sciencedirect.com/science/article/pii/S0038092X17308307.

[36] Carlo Brancucci Martinez-Anido, Benjamin Botor, Anthony R. Florita, Caroline Draxl, Siyuan Lu, Hendrik F. Hamann, and Bri-Mathias Hodge. The value of day-ahead solar power forecasting improvement. *Solar Energy*, 129:192–203, May 2016. ISSN 0038-092X. , 10.1016/j.solener.2016.01.049. URL https://www.sciencedirect.com/science/article/pii/S0038092X16000736.

[37] Shadi Goodarzi, H. Niles Perera, and Derek Bunn. The impact of renewable energy forecast

errors on imbalance volumes and electricity spot prices. *Energy Policy*, 134:110827, November 2019. ISSN 0301-4215. , 10.1016/j.enpol.2019.06.035. URL https://www.sciencedirect.com/science/article/pii/S0301421519304057.

[38] Jeff Tsao, Nate Lewis, and George Crabtree. Solar FAQs, 2006. URL https://old-www.sandia.gov/~jytsao/Solar%20FAQs.pdf.

[39] Franz Trieb, Hans Müller-Steinhagen, Jürgen Kern, Jürgen Scharfe, Malek Kabariti, and Ammar Al Taher. Technologies for large scale seawater desalination using concentrated solar radiation. *Desalination*, 235(1):33–43, January 2009. ISSN 0011-9164. , 10.1016/j.desal.2007.04.098. URL https://www.sciencedirect.com/science/article/pii/S0011916408005833.

[40] Ugo Pelay, Lingai Luo, Yilin Fan, Driss Stitou, and Mark Rood. Thermal energy storage systems for concentrated solar power plants. *Renewable and Sustainable Energy Reviews*, 79: 82–100, November 2017. ISSN 1364-0321. , 10.1016/j.rser.2017.03.139. URL https://www.sciencedirect.com/science/article/pii/S1364032117304021.

[41] Mohamed E. El-Khouly, Eithar El-Mohsnawy, and Shunichi Fukuzumi. Solar energy conversion: From natural to artificial photosynthesis. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, 31:36–83, June 2017. ISSN 1389-5567. , 10.1016/j.jphotochemrev.2017.02.001. URL https://www.sciencedirect.com/science/article/pii/S1389556716300727.

[42] IEA. Solar Energy: Mapping the Road Ahead. Technical report, IEA, Paris, 2019. URL https://www.iea.org/reports/solar-energy-mapping-the-road-ahead.

[43] RTE. Bilan prévisionnel long terme "Futurs énergétiques2050". Technical report, RTE, January 2021. URL https://assets.rte-france.com/prod/public/2021-01/Bilan%20Previsionnel%202050-consultation-complet.pdf.

[44] Cyril Voyant, Philippe Lauret, Gilles Notton, Jean-Laurent Duchaud, Alexis Fouilloy, Mathieu David, Zaher Mundher Yaseen, and Ted Soubdhan. A Monte Carlo Based Solar Radiation Forecastability Estimation. *Journal of Renewable and Sustainable Energy*, 2021. URL https://hal.archives-ouvertes.fr/hal-03162966.

[45] R. Tawn and J. Browell. A review of very short-term wind and solar power forecasting. *Renewable and Sustainable Energy Reviews*, 153:111758, January 2022. ISSN 13640321. , 10.1016/j.rser.2021.111758. URL https://linkinghub.elsevier.com/retrieve/pii/S1364032121010285.

[46] Maimouna Diagne, Mathieu David, Philippe Lauret, John Boland, and Nicolas Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27:65–76, November 2013. ISSN 1364-0321. , 10.1016/j.rser.2013.06.042. URL http://www.sciencedirect.com/science/article/pii/S1364032113004334.

[47] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, October 2016. ISSN 0038-092X. , 10.1016/j.solener.2016.06.069. URL http://www.sciencedirect.com/science/article/pii/S0038092X1630250X.

[48] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd. Rahim. Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management*, 156:459–497, January 2018. ISSN 0196-8904. , 10.1016/j.enconman.2017.11.019. URL http://www.sciencedirect.com/science/article/pii/S0196890417310622.

[49] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, July 2016. ISSN 0169-2070. , 10.1016/j.ijforecast.2016.02.001. URL http://www.sciencedirect.com/science/article/pii/S0169207016000133.

[50] Georges Kariniotakis. *Renewable Energy Forecasting: From Models to Applications*. Woodhead Publishing Series in Energy. Elsevier - Woodhead Publishing, 2017. URL https://hal-mines-paristech.archives-ouvertes.fr/hal-01542722.

[51] George Kariniotakis and Simon Camal. Improved weather modelling and forecasting dedicated to renewable energy applications. page 2, 2021. , https://doi.org/10.5194/egusphere-egu21-16219.

[52] G. Graditi, S. Ferlito, and G. Adinolfi. Comparison of Photovoltaic plant power production prediction methods using a large measured dataset. *Renewable Energy*, 90:513–519, May 2016. ISSN 0960-1481. , 10.1016/j.renene.2016.01.027. URL http://www.sciencedirect.com/science/article/pii/S0960148116300271.

[53] Niklas Benedikt Blum, Bijan Nouri, Stefan Wilbert, Thomas Schmidt, Ontje Lünsdorf, Jonas Stührenberg, Detlev Heinemann, Andreas Kazantzidis, and Robert Pitz-Paal. Cloud height measurement by a network of all-sky imagers. *Atmospheric Measurement Techniques*, 14(7): 5199–5224, July 2021. ISSN 1867-8548. , 10.5194/amt-14-5199-2021. URL https://amt.copernicus.org/articles/14/5199/2021/.

[54] Loïc Vallance. *Synergie des mesures pyranométriques et des images hémisphériques in-situ avec des images satellites météorologiques pour la prévision photovoltaïque*. phdthesis, PSL Research University, November 2018. URL https://pastel.archives-ouvertes.fr/tel-02097021.

[55] L. Mazorra Aguiar, B. Pereira, P. Lauret, F. Díaz, and M. David. Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. *Renewable Energy*, 97:599–610, November 2016. ISSN 0960-1481. , 10.1016/j.renene.2016.06.018. URL http://www.sciencedirect.com/science/article/pii/S0960148116305390.

[56] Annette Eschenbach, Guillermo Yepes, Christian Tenllado, José I. Gómez-Pérez, Luis Piñuel, Luis F. Zarzalejo, and Stefan Wilbert. Spatio-Temporal Resolution of Irradiance Samples in Machine Learning Approaches for Irradiance Forecasting. *IEEE Access*, 8:51518–51531, 2020. ISSN 2169-3536. , 10.1109/ACCESS.2020.2980775.

[57] R. J. Bessa, A. Trindade, Cátia S. P. Silva, and V. Miranda. Probabilistic solar power forecasting in smart grids using distributed information. *International Journal of Electrical Power & Energy Systems*, 72:16–23, November 2015. ISSN 0142-0615. , 10.1016/j.ijepes.2015.02.006. URL http://www.sciencedirect.com/science/article/pii/S0142061515000897.

[58] X. G. Agoua, R. Girard, and G. Kariniotakis. Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production. *IEEE Transactions on Sustainable Energy*, 9(2):538–546, April 2018. ISSN 1949-3029. , 10.1109/TSTE.2017.2747765.

[59] Anastasios Golnas, Joseph Bryan, Robert Wimbrow, Clifford Hansen, and Steve Voss. Performance assessment without pyranometers: Predicting energy output based on historical correlation. In *2011 37th IEEE Photovoltaic Specialists Conference*, pages 002006–002010, June 2011. , 10.1109/PVSC.2011.6186347.

[60] George Kariniotakis, Simon Camal, Ricardo Bessa, Pierre Pinson, Gregor Giebel, Quentin Libois, Raphaël Legrand, Matthias Lange, Stefan Wilbert, Bijan Nouri, Alexandre Neto,

Remco Verzijlbergh, Ganesh Sauba, George Sideratos, Efrosyni Korka, and Stephanie Petit. Smart4RES: Towards next generation forecasting tools of renewable energy production. other, oral, March 2020. URL https://meetingorganizer.copernicus.org/EGU2020/EGU2020-20205.html.

[61] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, October 2009. ISSN 0038092X. , 10.1016/j.solener.2009.05.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X09001364.

[62] R. Amaro e Silva, S. E. Haupt, and M. C. Brito. A regime-based approach for integrating wind information in spatio-temporal solar forecasting models. *Journal of Renewable and Sustainable Energy*, 11(5):056102, September 2019. ISSN 1941-7012. , 10.1063/1.5098763. URL http://aip.scitation.org/doi/10.1063/1.5098763.

[63] Changsong Chen, Shanxu Duan, Tao Cai, and Bangyin Liu. Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy*, 85(11):2856–2870, November 2011. ISSN 0038-092X. , 10.1016/j.solener.2011.08.027. URL http://www.sciencedirect.com/science/article/pii/S0038092X11003008.

[64] Jie Shi, Wei-Jen Lee, Yongqian Liu, Yongping Yang, and Peng Wang. Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. *IEEE Transactions on Industry Applications*, 48(3):1064–1069, May 2012. ISSN 1939-9367. , 10.1109/TIA.2012.2190816.

[65] Fei Wang, Zhao Zhen, Zengqiang Mi, Hongbin Sun, Shi Su, and Guang Yang. Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. *Energy and Buildings*, 86:427–438, January 2015. ISSN 0378-7788. , 10.1016/j.enbuild.2014.10.002. URL http://www.sciencedirect.com/science/article/pii/S0378778814008226.

[66] Marcelo Pinho Almeida, Oscar Perpiñán, and Luis Narvarte. PV power forecast using a nonparametric PV model. *Solar Energy*, 115:354–368, May 2015. ISSN 0038-092X. , 10.1016/j.solener.2015.03.006. URL http://www.sciencedirect.com/science/article/pii/S0038092X15001218.

[67] Thomas Carriere, Christophe Vernay, Sebastien Pitaval, and George Kariniotakis. A Novel Approach for Seamless Probabilistic Photovoltaic Power Forecasting Covering Multiple Time Frames. *IEEE Transactions on Smart Grid*, pages 1–1, 2019. ISSN 1949-3061. , 10.1109/TSG.2019.2951288.

[68] Aurélien Ben Daoud, Eric Sauquet, Guillaume Bontron, Charles Obled, and Michel Lang. Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. *Atmospheric Research*, 169:147–159, March 2016. ISSN 01698095. , 10.1016/j.atmosres.2015.09.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0169809515002951.

[69] Mohammed Jasim M. Al Essa. Power management of grid-integrated energy storage batteries with intermittent renewables. *Journal of Energy Storage*, 31:101762, October 2020. ISSN 2352152X. , 10.1016/j.est.2020.101762. URL https://linkinghub.elsevier.com/retrieve/pii/S2352152X20315991.

[70] Matheus Sabino Viana, Giovanni Manassero, and Miguel E.M. Udaeta. Analysis of demand response and photovoltaic distributed generation as resources for power utility planning. *Applied Energy*, 217:456–466, May 2018. ISSN 03062619. , 10.1016/j.apenergy.2018.02.153. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261918302873.

[71] Kevin Bellinguer, Robin Girard, Guillaume Bontron, and Georges Kariniotakis. Short-Term Photovoltaic Generation Forecasting Enhanced by Satellite Derived Irradiance. In *26th International Conference & Exhibition on Electricity Distribution (CIRED 2021)*, pages 1–6, Virtual Event, Switzerland, September 2021. CIRED. URL https://hal.archives-ouvertes.fr/hal-03407898.

[72] K. Bellinguer, R. Girard, G. Bontron, and G. Kariniotakis. Short-term Forecasting of Photovoltaic Generation based on Conditioned Learning of Geopotential Fields. In *2020 55th International Universities Power Engineering Conference (UPEC)*, pages 1–6, September 2020. , 10.1109/UPEC49904.2020.9209858.

[73] Alfonso Delgado-Bonal and Alexander Marshak. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy*, 21(6):541, June 2019. , 10.3390/e21060541. URL https://www.mdpi.com/1099-4300/21/6/541.

[74] Philippe Lauret, Mathieu David, and Pierre Pinson. Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254–271, December 2019. , 10.1016/j.solener.2019.10.041. URL https://hal.archives-ouvertes.fr/hal-02351342.

[75] Dazhi Yang. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). *Journal of Renewable and Sustainable Energy*, 11(2):022701, March 2019. ISSN 1941-7012. , 10.1063/1.5087462. URL http://aip.scitation.org/doi/10.1063/1.5087462.

[76] D. W. van der Meer, J. Widén, and J. Munkhammar. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81:1484–1512, January 2018. ISSN 1364-0321. , 10.1016/j.rser.2017.05.212. URL http://www.sciencedirect.com/science/article/pii/S1364032117308523.

[77] Mellit, Massi Pavan, Ogliari, Leva, and Lughi. Advanced Methods for Photovoltaic Output Power Forecasting: A Review. *Applied Sciences*, 10(2):487, January 2020. ISSN 2076-3417. , 10.3390/app10020487. URL https://www.mdpi.com/2076-3417/10/2/487.

[78] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

[79] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., San Francisco, CA, USA, 1990. ISBN 978-0-8162-1104-3.

[80] Dazhi Yang, Jan Kleissl, Christian A. Gueymard, Hugo T. C. Pedro, and Carlos F. M. Coimbra. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, 168:60–101, July 2018. ISSN 0038-092X. , 10.1016/j.solener.2017.11.023. URL https://www.sciencedirect.com/science/article/pii/S0038092X17310022.

[81] Chen Yang. A novel ARX-based multi-scale spatio-temporal solar power forecast model. In *2012 North American Power Symposium (NAPS)*, pages 1–6, September 2012. , 10.1109/NAPS.2012.6336383.

[82] John Boland. Spatial-temporal forecasting of solar radiation. *Renewable Energy*, 75:607–616, March 2015. ISSN 0960-1481. , 10.1016/j.renene.2014.10.035. URL http://www.sciencedirect.com/science/article/pii/S0960148114006624.

[83] R. J. Bessa, A. Trindade, and V. Miranda. Spatial-Temporal Solar Power Forecasting for Smart Grids. *IEEE Transactions on Industrial Informatics*, 11(1):232–241, February 2015. ISSN 1551-3203. , 10.1109/TII.2014.2365703.

[84] Maïna André, Sophie Dabo-Niang, Ted Soubdhan, and Hanany Ould-Baba. Predictive spatio-temporal model for spatially sparse global solar radiation data. *Energy*, 111:599–608, September 2016. ISSN 03605442. , 10.1016/j.energy.2016.06.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0360544216307769.

[85] Dazhi Yang, Zibo Dong, Thomas Reindl, Panida Jirutitijaroen, and Wilfred M. Walsh. Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. *Solar Energy*, 103:550–562, May 2014. ISSN 0038-092X. , 10.1016/j.solener.2014.01.024. URL http://www.sciencedirect.com/science/article/pii/S0038092X14000425.

[86] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. URL https://www.jstor.org/stable/2346178.

[87] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33:1–22, February 2010. ISSN 1548-7660. , 10.18637/jss.v033.i01. URL https://doi.org/10.18637/jss.v033.i01.

[88] Tim Januschowski, Yuyang Wang, Kari Torkkola, Timo Erkkilä, Hilaf Hasson, and Jan Gasthaus. Forecasting with trees. *International Journal of Forecasting*, page S0169207021001679, December 2021. ISSN 01692070. , 10.1016/j.ijforecast.2021.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207021001679.

[89] Kevin Bellinguer, Valentin Mahler, Simon Camal, and Georges Kariniotakis. Probabilistic Forecasting of Regional Wind Power Generation for the EEM20 Competition: a Physics-oriented Machine Learning Approach. In *2020 17th International Conference on the European Energy Market (EEM)*, pages 1–6, September 2020. , 10.1109/EEM49802.2020.9221960.

[90] Da Liu and Kun Sun. Random forest solar power forecast based on classification optimization. *Energy*, 187:115940, November 2019. ISSN 0360-5442. , 10.1016/j.energy.2019.115940. URL http://www.sciencedirect.com/science/article/pii/S036054421931624X.

[91] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. , 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[92] Erwan Scornet. Trees, forests, and impurity-based variable importance. *arXiv:2001.04295 [math, stat]*, November 2020. URL http://arxiv.org/abs/2001.04295.

[93] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. PhD thesis, arXiv, June 2015. URL http://arxiv.org/abs/1407.7502. arXiv:1407.7502 [stat].

[94] Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77:1–17, March 2017. ISSN 1548-7660. , 10.18637/jss.v077.i01. URL https://doi.org/10.18637/jss.v077.i01.

[95] Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Siyuan Lu, Hendrik F. Hamann, Venkat Banunarayanan, and Anna M. Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, January 2015. ISSN 0038092X. , 10.1016/j.solener.2014.10.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X14005027.

[96] Loïc Vallance, Bruno Charbonnier, Nicolas Paul, Stéphanie Dubost, and Philippe Blanc. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. *Solar Energy*, 150:408–422, 2017. ISSN 0038-092X. , 10.1016/j.solener.2017.04.064. URL http://www.sciencedirect.com/science/article/pii/S0038092X17303687.

[97] Dazhi Yang, Stefano Alessandrini, Javier Antonanzas, Fernando Antonanzas-Torres, Viorel Badescu, Hans Georg Beyer, Robert Blaga, John Boland, Jamie M. Bright, Carlos F.M. Coimbra, Mathieu David, Azeddine Frimane, Christian A. Gueymard, Tao Hong, Merlinde J. Kay, Sven Killinger, Jan Kleissl, Philippe Lauret, Elke Lorenz, Dennis van der Meer, Marius Paulescu, Richard Perez, Oscar Perpinan-Lamigueiro, Ian Marius Peters, Gordon Reikard, David Renné, Yves-Marie Saint-Drenan, Yong Shuai, Ruben Urraca, Hadrien Verbois, Frank Vignola, Cyril Voyant, and Jie Zhang. Verification of deterministic solar forecasts. *Solar Energy*, 210:20–37, November 2020. ISSN 0038092X. , 10.1016/j.solener.2020.04.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X20303947.

[98] Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef, Moh Yamani Idna Idris, Willem Van Deventer, Bend Horan, and Alex Stojcevski. Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81:912–928, January 2018. ISSN 1364-0321. , 10.1016/j.rser.2017.08.017. URL http://www.sciencedirect.com/science/article/pii/S1364032117311620.

[99] Francis X. Diebold. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1, January 2015. ISSN 0735-0015. , 10.1080/07350015.2014.983236. URL https://doi.org/10.1080/07350015.2014.983236.

[100] Xwegnon Agoua. *Développement de méthodes spatio-temporelles pour la prévision à court terme de la production photovoltaïque*. PhD thesis, Mines ParisTech, December 2017. URL https://pastel.archives-ouvertes.fr/tel-01878943/document.

[101] Song Chen, Peng Li, David Brady, and Brad Lehman. Determining the optimum grid-connected photovoltaic inverter size. *Solar Energy*, 87:96–116, January 2013. ISSN 0038-092X. , 10.1016/j.solener.2012.09.012. URL https://www.sciencedirect.com/science/article/pii/S0038092X12003362.

[102] Reinhard Mackensen, Yves-Marie Saint-Drenan, Dominik Jost, Rafael Fritz, Nazgul Asanalieva, Martin Widdel, and Markus Hahler. Regelenergie Durch Wind- und Photovoltaikparks. Technical report, Fraunhofer, Kassel, July 2017. URL https://www.iee.fraunhofer.de/content/dam/iee/energiesystemtechnik/de/Dokumente/Projekte/20170814_ReWP_Abschluss_final.pdf.

[103] Javier Marcos, Luis Marroyo, Eduardo Lorenzo, David Alvira, and Eloisa Izco. Power output fluctuations in large scale pv plants: One year observations with one second resolution and a derived analytic model. *Progress in Photovoltaics: Research and Applications*, 19(2):218–227, 2011. ISSN 1099-159X. , 10.1002/pip.1016. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.1016.

[104] T. Kato, Y. Manabe, T. Funabashi, K. Yoshiura, M. Kurimoto, and Y. Suzuoki. A study on several hours ahead forecasting of spatial average irradiance using NWP model and satellite infrared image. In *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–8, October 2016. , 10.1109/PMAPS.2016.7764096.

[105] Benoît Gschwind, Lionel Ménard, Michel Albuisson, and Lucien Wald. Converting a successful research project into a sustainable service: The case of the SoDa Web service. *Environmental Modelling & Software*, 21(11):1555–1561, November 2006. ISSN 1364-8152. , 10.1016/j.envsoft.2006.05.002. URL https://www.sciencedirect.com/science/article/pii/S1364815206001137.

[106] Philippe Blanc, Benoît Gschwind, Mireille Lefèvre, and Lucien Wald. The HelioClim Project:

Surface Solar Irradiance Data for Climate Applications. *Remote Sensing*, 3(2):343–361, February 2011. , 10.3390/rs3020343. URL https://hal-mines-paristech.archives-ouvertes.fr/hal-00566995/document.

[107] José A. Ruiz-Arias and Christian A. Gueymard. Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface. *Solar Energy*, 168:10–29, July 2018. ISSN 0038-092X. , 10.1016/j.solener.2018.02.008. URL https://www.sciencedirect.com/science/article/pii/S0038092X18301257.

[108] EUMETSAT. Spinning Enhanced Visible and InfraRed Imager (SEVIRI) | EUMETSAT, May 2020. URL https://www.eumetsat.int/seviri.

[109] Meteo-France. Les principaux satellites utilisés à Météo-France. URL http://www.meteo-spatiale.fr/content/perenne/cours/05-LE-MOAL-satellites-utilises-a-meteo-france.pdf.

[110] Hervé Le Gléau, Marcel Derrien, and D Lannion. L'observation des nuages et de leurs propriétés physiques par satellite. Technical report, Meteo France, 2010.

[111] Maximilien PATOU. *Analyse temporelle des propriétés optiques, microphysiques et macrophysiques de systèmes nuageux fortement précipitants à partir de SEVIRI/MSG*. PhD thesis, Université de Lille, 2018. URL https://www-loa.univ-lille1.fr/documents/LOA/formation/theses/2018_Patou.pdf.

[112] ECMWF. IFS Documentation CY47R1. Technical report, ECMWF, 2020.

[113] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone. An Analog Ensemble for Short-term Probabilistic Solar Power Forecast. *Applied Energy*, 157:95–110, November 2015. ISSN 0306-2619. , 10.1016/j.apenergy.2015.08.011. URL http://www.sciencedirect.com/science/article/pii/S0306261915009368.

[114] Annette Hammer, Jan Kühnert, Kailash Weinreich, and Elke Lorenz. Short-Term Forecasting of Surface Solar Irradiance Based on Meteosat-SEVIRI Data Using a Nighttime Cloud Index. *Remote Sensing*, 7(7):9070–9090, July 2015. ISSN 2072-4292. , 10.3390/rs70709070. URL http://www.mdpi.com/2072-4292/7/7/9070.

[115] Martin János Mayer and Gyula Gróf. Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, 283:116239, February 2021. ISSN 0306-2619. , 10.1016/j.apenergy.2020.116239. URL https://www.sciencedirect.com/science/article/pii/S0306261920316330.

[116] Lucien Wald. BASICS IN SOLAR RADIATION AT EARTH SURFACE, January 2018. URL https://hal-mines-paristech.archives-ouvertes.fr/hal-01676634/document.

[117] Soda-Pro. CAMS McClear - www.soda-pro.com. URL http://www.soda-pro.com/web-services/radiation/cams-mcclear.

[118] Robin Hogan. Radiation Quantities in the ECMWF model and MARS. page 9, 2015. URL https://www.ecmwf.int/node/18490.

[119] Christian A. Gueymard and Jose A. Ruiz-Arias. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. *Solar Energy*, 128:1–30, April 2016. ISSN 0038-092X. , 10.1016/j.solener.2015.10.010. URL https://www.sciencedirect.com/science/article/pii/S0038092X15005435.

[120] N. A. Engerer. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. *Solar Energy*, 116:215–237, June 2015. ISSN 0038-092X. ,

10.1016/j.solener.2015.04.012. URL https://www.sciencedirect.com/science/article/pii/S0038092X15001905.

[121] Markku Järvelä, Kari Lappalainen, and Seppo Valkealahti. Characteristics of the cloud enhancement phenomenon and PV power plants. *Solar Energy*, 196:137–145, January 2020. ISSN 0038-092X. , 10.1016/j.solener.2019.11.090. URL https://www.sciencedirect.com/science/article/pii/S0038092X19311909.

[122] Barbara Ridley, John Boland, and Philippe Lauret. Modelling of diffuse solar fraction with multiple predictors. *Renewable Energy*, 35(2):478–483, February 2010. ISSN 09601481. , 10.1016/j.renene.2009.07.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0960148109003012.

[123] Redlich García Rojas, Natalia Alvarado, John Boland, Rodrigo Escobar, and Armando Castillejo-Cuberos. Diffuse fraction estimation using the BRL model and relationship of predictors under Chilean, Costa Rican and Australian climatic conditions. *Renewable Energy*, 136:1091–1106, June 2019. ISSN 0960-1481. , 10.1016/j.renene.2018.09.079. URL https://www.sciencedirect.com/science/article/pii/S0960148118311534.

[124] Leonardo F. L. Lemos, Allan R. Starke, John Boland, José M. Cardemil, Rubinei D. Machado, and Sergio Colle. Assessment of solar radiation components in Brazil using the BRL model. *Renewable Energy*, 108:569–580, August 2017. ISSN 0960-1481. , 10.1016/j.renene.2017.02.077. URL https://www.sciencedirect.com/science/article/pii/S0960148117301635.

[125] Makbul A. M. Ramli, Ssennoga Twaha, and Yusuf A. Al-Turki. Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. *Energy Conversion and Management*, 105:442–452, November 2015. ISSN 0196-8904. , 10.1016/j.enconman.2015.07.083. URL https://www.sciencedirect.com/science/article/pii/S0196890415007426.

[126] Dazhi Yang. Solar radiation on inclined surfaces: Corrections and benchmarks. *Solar Energy*, 136:288–302, October 2016. ISSN 0038-092X. , 10.1016/j.solener.2016.06.062. URL http://www.sciencedirect.com/science/article/pii/S0038092X16302432.

[127] Nóra Varga and Martin János Mayer. Model-based analysis of shading losses in ground-mounted photovoltaic power plants. *Solar Energy*, 216:428–438, March 2021. ISSN 0038-092X. , 10.1016/j.solener.2021.01.047. URL https://www.sciencedirect.com/science/article/pii/S0038092X21000633.

[128] Richard Perez, Pierre Ineichen, Robert Seals, Joseph Michalsky, and Ronald Stewart. Modeling daylight availability and irradiance components from direct and global irradiance. *Solar Energy*, 44(5):271–289, January 1990. ISSN 0038-092X. , 10.1016/0038-092X(90)90055-H. URL http://www.sciencedirect.com/science/article/pii/0038092X9090055H.

[129] Fritz Kasten and Andrew T. Young. Revised optical air mass tables and approximation formula. *Applied Optics*, 28(22):4735, November 1989. ISSN 0003-6935, 1539-4522. , 10.1364/AO.28.004735. URL https://www.osapublishing.org/abstract.cfm?URI=ao-28-22-4735.

[130] John A Duffie and William A Beckman. *Solar Engineering of Thermal Processes*. John Wiley & Sons, 2013. ISBN 0-470-87366-3.

[131] Colienne Demain, Michel Journée, and Cédric Bertrand. Evaluation of different models to estimate the global solar radiation on inclined surfaces. *Renewable Energy*, 50:710–721, February 2013. ISSN 0960-1481. , 10.1016/j.renene.2012.07.031. URL https://www.sciencedirect.com/science/article/pii/S0960148112004570.

[132] Christian A. Gueymard. Direct and indirect uncertainties in the prediction of tilted irradiance for solar engineering applications. *Solar Energy*, 83(3):432–444, March 2009. ISSN 0038-092X. , 10.1016/j.solener.2008.11.004. URL https://www.sciencedirect.com/science/article/pii/S0038092X08002983.

[133] Yves-Marie Saint-Drenan and Thibaut Barbier. Data-analysis and modelling of the effect of inter-row shading on the power production of photovoltaic plants. *Solar Energy*, 184:127–147, May 2019. ISSN 0038-092X. , 10.1016/j.solener.2019.03.086. URL https://www.sciencedirect.com/science/article/pii/S0038092X19303147.

[134] Martin János Mayer and Gyula Gróf. Techno-economic optimization of grid-connected, ground-mounted photovoltaic power plants by genetic algorithm based on a comprehensive mathematical model. *Solar Energy*, 202:210–226, May 2020. ISSN 0038-092X. , 10.1016/j.solener.2020.03.109. URL https://www.sciencedirect.com/science/article/pii/S0038092X20303558.

[135] Anurag Singh Yadav and V. Mukherjee. Conventional and advanced PV array configurations to extract maximum power under partial shading conditions: A review. *Renewable Energy*, 178:977–1005, November 2021. ISSN 0960-1481. , 10.1016/j.renene.2021.06.029. URL https://www.sciencedirect.com/science/article/pii/S0960148121008958.

[136] Jirada Gosumbonggot and Goro Fujita. Global Maximum Power Point Tracking under Shading Condition and Hotspot Detection Algorithms for Photovoltaic Systems. *Energies*, 12(5):882, January 2019. , 10.3390/en12050882. URL https://www.mdpi.com/1996-1073/12/5/882.

[137] Shifeng Deng, Zhen Zhang, Chenhui Ju, Jingbing Dong, Zhengyue Xia, Xinchun Yan, Tao Xu, and Guoqiang Xing. Research on hot spot risk for high-efficiency solar module. *Energy Procedia*, 130:77–86, September 2017. ISSN 18766102. , 10.1016/j.egypro.2017.09.399. URL https://linkinghub.elsevier.com/retrieve/pii/S1876610217344909.

[138] Assaf Peled and Joseph Appelbaum. The view-factor effect shaping of I-V characteristics. *Progress in Photovoltaics: Research and Applications*, 26(4):273–280, 2018. ISSN 1099-159X. , 10.1002/pip.2979. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2979.

[139] Avi Aronescu and Joseph Appelbaum. The Effect of Collector Shading and Masking on Optimized PV Field Designs. *Energies*, 12(18):3471, January 2019. , 10.3390/en12183471. URL https://www.mdpi.com/1996-1073/12/18/3471.

[140] Ruben URRACA VALLE. Horizon profile, October 2019. URL https://ec.europa.eu/jrc/en/PVGIS/tools/horizon.

[141] Klemens Ilse, Leonardo Micheli, Benjamin W. Figgis, Katja Lange, David Daßler, Hamed Hanifi, Fabian Wolfertstetter, Volker Naumann, Christian Hagendorf, Ralph Gottschalg, and Jörg Bagdahn. Techno-Economic Assessment of Soiling Losses and Mitigation Strategies for Solar Power Generation. *Joule*, 3(10):2303–2321, October 2019. ISSN 2542-4351. , 10.1016/j.joule.2019.08.019. URL https://www.sciencedirect.com/science/article/pii/S2542435119304222.

[142] A. Younis and Y. Alhorr. Modeling of dust soiling effects on solar photovoltaic performance: A review. *Solar Energy*, 220:1074–1088, May 2021. ISSN 0038-092X. , 10.1016/j.solener.2021.04.011. URL https://www.sciencedirect.com/science/article/pii/S0038092X21002929.

[143] Reinhart Appels, Buvaneshwari Lefevre, Bert Herteleer, Hans Goverde, Alexander Beerten, Robin Paesen, Klaas De Medts, Johan Driesen, and Jef Poortmans. Effect of soiling on photovoltaic modules. *Solar Energy*, 96:283–291, October 2013. ISSN 0038-092X. , 10.1016/j.solener.2013.07.017. URL https://www.sciencedirect.com/science/article/pii/S0038092X1300282X.

[144] Leonardo Micheli, Eduardo F. Fernández, Jorge T. Aguilera, and Florencia Almonacid. Economics of seasonal photovoltaic soiling and cleaning optimization scenarios. *Energy*, 215: 119018, January 2021. ISSN 0360-5442. , 10.1016/j.energy.2020.119018. URL https://www.sciencedirect.com/science/article/pii/S0360544220321253.

[145] Michael G. Deceglie, Leonardo Micheli, and Matthew Muller. Quantifying Soiling Loss Directly From PV Yield. *IEEE Journal of Photovoltaics*, 8(2):547–551, March 2018. ISSN 2156-3403. , 10.1109/JPHOTOV.2017.2784682.

[146] Nicole Lindsay. Implementation of a photovoltaic power diagnostic tool in the Meso-NH atmospheric model. Technical report, Ecole Polytechnique Fédérale de Lausanne, 2018. URL https://www.umr-cnrm.fr/IMG/pdf/pdm_rapport_final.pdf.

[147] N. Martin and J. M. Ruiz. Calculation of the PV modules angular losses under field conditions by means of an analytical model. *Solar Energy Materials and Solar Cells*, 70(1): 25–38, December 2001. ISSN 0927-0248. , 10.1016/S0927-0248(00)00408-6. URL https://www.sciencedirect.com/science/article/pii/S0927024800004086.

[148] N. Lindsay, Q. Libois, J. Badosa, A. Migan-Dubois, and V. Bourdin. Errors in PV power modelling due to the lack of spectral and angular details of solar irradiance inputs. *Solar Energy*, 197:266–278, February 2020. ISSN 0038-092X. , 10.1016/j.solener.2019.12.042. URL https://www.sciencedirect.com/science/article/pii/S0038092X19312563.

[149] W. De Soto, S. A. Klein, and W. A. Beckman. Improvement and validation of a model for photovoltaic array performance. *Solar Energy*, 80(1):78–88, January 2006. ISSN 0038-092X. , 10.1016/j.solener.2005.06.010. URL https://www.sciencedirect.com/science/article/pii/S0038092X05002410.

[150] N. Martin and J. M. Ruiz. Corrigendum to "Calculation of the PV modules angular losses under field conditions by means of an analytical model" [Sol. Energy Mater. Sol. Cells 70 (1) (2001) 25–38]. *Solar Energy Materials and Solar Cells*, 110:154, March 2013. ISSN 0927-0248. , 10.1016/j.solmat.2012.11.002. URL https://www.sciencedirect.com/science/article/pii/S0927024812004990.

[151] A. Dobos. PVWatts Version 5 Manual. Technical Report NREL/TP-6A20-62641, 1158421, PVWatts, September 2014. URL http://www.osti.gov/servlets/purl/1158421/.

[152] Soteris A. Kalogirou. Photovoltaic Systems. In *Solar Energy Engineering*, pages 481–540. Elsevier, 2014. ISBN 978-0-12-397270-5. URL https://linkinghub.elsevier.com/retrieve/pii/B9780123972705000091.

[153] B. Kumaragurubaran and S. Anandhi. Reduction of reflection losses in solar cell using Anti Reflective coating. In *2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, pages 155–157, April 2014. , 10.1109/ICCPEIC.2014.6915357.

[154] Antonin Faes, Jacques Levrat, Jonathan Champliaud, Matthieu Despeisse, Mauro Caccivio, and Brian Custodio. Incident angle modifier evaluation for DSM coating technologies, 2019.

[155] Bill Marion. Numerical method for angle-of-incidence correction factors for diffuse radiation incident photovoltaic modules. *Solar Energy*, 147:344–348, May 2017. ISSN 0038-092X. , 10.1016/j.solener.2017.03.027. URL https://www.sciencedirect.com/science/article/pii/S0038092X17301883.

[156] M. J. Brandemuehl and W. A. Beckman. Transmission of diffuse radiation through CPC and flat plate collector glazings. *Solar Energy*, 24(5):511–513, January 1980. ISSN 0038-092X. ,

10.1016/0038-092X(80)90320-5. URL https://www.sciencedirect.com/science/article/pii/0038092X80903205.

[157] Mitchell Lee and Alex Panchula. Spectral correction for photovoltaic module performance based on air mass and precipitable water. In *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, pages 1351–1356, June 2016. , 10.1109/PVSC.2016.7749836.

[158] Sophie Pelland, Colin Beswick, Didier Thevenard, Alexandre Côté, Abhijeet Pai, and Yves Poissant. Development and Testing of the PVSPEC Model of Photovoltaic Spectral Mismatch Factor. In *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, pages 1258–1264, June 2020. , 10.1109/PVSC45281.2020.9300932.

[159] Álvaro Fernández-Solas, Leonardo Micheli, Florencia Almonacid, and Eduardo F. Fernández. Optical degradation impact on the spectral performance of photovoltaic technology. *Renewable and Sustainable Energy Reviews*, 141:110782, May 2021. ISSN 1364-0321. , 10.1016/j.rser.2021.110782. URL https://www.sciencedirect.com/science/article/pii/S1364032121000770.

[160] Yujuan He, Jie Liu, Shi-Joon Sung, and Chih-hung Chang. Downshifting and antireflective thin films for solar module power enhancement. *Materials & Design*, 201:109454, March 2021. ISSN 0264-1275. , 10.1016/j.matdes.2021.109454. URL https://www.sciencedirect.com/science/article/pii/S0264127521000071.

[161] Evaldo C. Gouvêa, Pedro M. Sobrinho, and Teófilo M. Souza. Spectral Response of Polycrystalline Silicon Photovoltaic Cells under Real-Use Conditions. *Energies*, 10(8):1178, August 2017. , 10.3390/en10081178. URL https://www.mdpi.com/1996-1073/10/8/1178.

[162] Thomas Huld, Gabi Friesen, Artur Skoczek, Robert P. Kenny, Tony Sample, Michael Field, and Ewan D. Dunlop. A power-rating model for crystalline silicon PV modules. *Solar Energy Materials and Solar Cells*, 95(12):3359–3369, December 2011. ISSN 0927-0248. , 10.1016/j.solmat.2011.07.026. URL https://www.sciencedirect.com/science/article/pii/S0927024811004442.

[163] Timothy J Silverman, Ulrike Jahn, Gabi Friesen, International Energy Agency, and Photovoltaic Power Systems Programme. *Characterisation of performance of thin-film photovoltaic technologies IEA PVPS task 13, subtask 3.1*. IAE, 2014. ISBN 978-3-906042-17-6. URL https://edocs.tib.eu/files/e01fb16/856981168.pdf.

[164] Justo José Roberts, Andrés A. Mendiburu Zevallos, and Agnelo Marotta Cassula. Assessment of photovoltaic performance models for system simulation. *Renewable and Sustainable Energy Reviews*, 72:1104–1123, May 2017. ISSN 1364-0321. , 10.1016/j.rser.2016.10.022. URL https://www.sciencedirect.com/science/article/pii/S1364032116306712.

[165] B. Marion. Comparison of predictive models for photovoltaic module performance. In *2008 33rd IEEE Photovoltaic Specialists Conference*, pages 1–6, May 2008. , 10.1109/PVSC.2008.4922586.

[166] E. Skoplaki and J. A. Palyvos. On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Solar Energy*, 83(5):614–624, May 2009. ISSN 0038-092X. , 10.1016/j.solener.2008.10.008. URL https://www.sciencedirect.com/science/article/pii/S0038092X08002788.

[167] Swapnil Dubey, Jatin Narotam Sarvaiya, and Bharath Seshadri. Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World – A Review. *Energy Procedia*, 33:311–321, January 2013. ISSN 1876-6102. , 10.1016/j.egypro.2013.05.072. URL https://www.sciencedirect.com/science/article/pii/S1876610213000829.

[168] J. Ascencio-Vásquez, J. Bevc, K. Reba, K. Brecl, M. Jankovec, and M. Topič. Advanced PV performance modelling based on different levels of irradiance data accuracy. *Energies*, 13(9), 2020. , 10.3390/en13092166.

[169] P. Rodrigo, C. Rus, F. Almonacid, P. J. Pérez-Higueras, and G. Almonacid. A new method for estimating angular, spectral and low irradiance losses in photovoltaic systems using an artificial neural network model in combination with the Osterwald model. *Solar Energy Materials and Solar Cells*, 96:186–194, January 2012. ISSN 0927-0248. , 10.1016/j.solmat.2011.09.054. URL https://www.sciencedirect.com/science/article/pii/S092702481100540X.

[170] I. de la Parra, M. Muñoz, E. Lorenzo, M. García, J. Marcos, and F. Martínez-Moreno. PV performance modelling: A review in the light of quality assurance for large PV plants. *Renewable and Sustainable Energy Reviews*, 78:780–797, October 2017. ISSN 1364-0321. , 10.1016/j.rser.2017.04.080. URL https://www.sciencedirect.com/science/article/pii/S1364032117305920.

[171] A. Hadj Arab, B. Taghezouit, K. Abdeladim, S. Semaoui, A. Razagui, A. Gherbi, S. Boulahchiche, and I. Hadj Mahammed. Maximum power output performance modeling of solar photovoltaic modules. *Energy Reports*, 6:680–686, February 2020. ISSN 2352-4847. , 10.1016/j.egyr.2019.09.049. URL https://www.sciencedirect.com/science/article/pii/S235248471930561X.

[172] Fabrizio Sossan, Enrica Scolari, Rahul Gupta, and Mario Paolone. Solar irradiance estimations for modeling the variability of photovoltaic generation and assessing violations of grid constraints: A comparison between satellite and pyranometers measurements with load flow simulations. *Journal of Renewable and Sustainable Energy*, 11(5):056103, September 2019. , 10.1063/1.5109076. URL https://aip.scitation.org/doi/10.1063/1.5109076.

[173] Wilhelm Durisch, Bernd Bitnar, Jean-C. Mayor, Helmut Kiess, King-hang Lam, and Josie Close. Efficiency model for photovoltaic modules and demonstration of its application to energy yield estimation. *Solar Energy Materials and Solar Cells*, 91(1):79–84, January 2007. ISSN 0927-0248. , 10.1016/j.solmat.2006.05.011. URL https://www.sciencedirect.com/science/article/pii/S0927024806003345.

[174] F. Mavromatakis, F. Vignola, and B. Marion. Low irradiance losses of photovoltaic modules. *Solar Energy*, 157:496–506, November 2017. ISSN 0038-092X. , 10.1016/j.solener.2017.08.062. URL https://www.sciencedirect.com/science/article/pii/S0038092X17307430.

[175] Bill Marion, Michael G. Deceglie, and Timothy J. Silverman. Analysis of measured photovoltaic module performance for Florida, Oregon, and Colorado locations. *Solar Energy*, 110:736–744, December 2014. ISSN 0038-092X. , 10.1016/j.solener.2014.10.017. URL https://www.sciencedirect.com/science/article/pii/S0038092X14005039.

[176] Blaz Kirn, Kristijan Brecl, and Marko Topic. A new PV module performance model based on separation of diffuse and direct light. *Solar Energy*, 113:212–220, March 2015. ISSN 0038-092X. , 10.1016/j.solener.2014.12.029. URL https://www.sciencedirect.com/science/article/pii/S0038092X14006203.

[177] Martin Libra, Vladislav Poulek, and Pavel Kouřím. Temperature changes of I-V characteristics of photovoltaic cells as a consequence of the Fermi energy level shift. 63:6, 2017. , 10.17221/38/2015-RAE.

[178] Michael Koehl, Markus Heck, Stefan Wiesmeier, and Jochen Wirth. Modeling of the nominal operating cell temperature based on outdoor weathering. *Solar Energy Materials and Solar*

*Cells*, 95(7):1638–1646, July 2011. ISSN 0927-0248. , 10.1016/j.solmat.2011.01.020. URL https://www.sciencedirect.com/science/article/pii/S0927024811000304.

[179] Sillia. Catalogue Sillia. Technical report, Sillia Energie, 2012. URL http://www.sillia.com/IMG/pdf/c0001_v5_catalogue_sillia_fr_pv_cycle_aqpv2012_bd.pdf.

[180] C. Schwingshackl, M. Petitta, J. E. Wagner, G. Belluardo, D. Moser, M. Castelli, M. Zebisch, and A. Tetzlaff. Wind Effect on PV Module Temperature: Analysis of Different Techniques for an Accurate Estimation. *Energy Procedia*, 40:77–86, January 2013. ISSN 1876-6102. , 10.1016/j.egypro.2013.08.010. URL https://www.sciencedirect.com/science/article/pii/S1876610213016044.

[181] M. Mattei, G. Notton, C. Cristofari, M. Muselli, and P. Poggi. Calculation of the polycrystalline PV module temperature using a simple method of energy balance. *Renewable Energy*, 31(4):553–567, April 2006. ISSN 0960-1481. , 10.1016/j.renene.2005.03.010. URL https://www.sciencedirect.com/science/article/pii/S096014810500073X.

[182] George Makrides, Bastian Zinsser, Matthew Norton, and George E. Georghiou. *Performance of Photovoltaics Under Actual Operating Conditions*. IntechOpen, March 2012. ISBN 978-953-51-0304-2. URL https://www.intechopen.com/chapters/32596.

[183] Michael G. Deceglie, Dirk C. Jordan, Ambarish Nag, Adam Shinn, and Chris Deline. Fleet-Scale Energy-Yield Degradation Analysis Applied to Hundreds of Residential and Nonresidential Photovoltaic Systems. *IEEE Journal of Photovoltaics*, 9(2):476–482, March 2019. ISSN 2156-3403. , 10.1109/JPHOTOV.2018.2884948.

[184] N.M. Pearsall. Introduction to photovoltaic system performance. In *The Performance of Photovoltaic (PV) Systems*, pages 1–19. Elsevier, 2017. ISBN 978-1-78242-336-2. URL https://linkinghub.elsevier.com/retrieve/pii/B978178242336200001X.

[185] Anton Driesse, Praveen Jain, and Steve Harrison. Beyond the curves: Modeling the electrical efficiency of photovoltaic inverters. In *2008 33rd IEEE Photovoltaic Specialists Conference*, pages 1–6, May 2008. , 10.1109/PVSC.2008.4922827.

[186] The world's most valuable resource is no longer oil, but data. *The Economist*, May 2017. ISSN 0013-0613. URL https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

[187] Will Murphy. Data is the New Oil, January 2022. URL https://towardsdatascience.com/data-is-the-new-oil-f11440e80dd0.

[188] Marie de Largentiere. Mairie > La Centrale Photovoltaïque • Commune de Largentière - Ardèche. URL https://largentiere.fr/mairie/la-centrale-photovoltaique/.

[189] Sven Killinger, Nicholas Engerer, and Björn Müller. QCPV: A quality control algorithm for distributed photovoltaic array power output. *Solar Energy*, 143:120–131, February 2017. ISSN 0038-092X. , 10.1016/j.solener.2016.12.053. URL https://www.sciencedirect.com/science/article/pii/S0038092X16306600.

[190] A. Mellit, G. M. Tina, and S. A. Kalogirou. Fault detection and diagnosis methods for photovoltaic systems: A review. *Renewable and Sustainable Energy Reviews*, 91:1–17, August 2018. ISSN 1364-0321. , 10.1016/j.rser.2018.03.062. URL https://www.sciencedirect.com/science/article/pii/S1364032118301370.

[191] Vincent P. Lonij, Adria E. Brooks, Kevin Koch, and Alexander D. Cronin. Analysis of 80 rooftop PV systems in the Tucson, AZ area. In *2012 38th IEEE Photovoltaic Specialists Conference*, pages 000549–000553, June 2012. , 10.1109/PVSC.2012.6317674.

[192] S. K. Firth, K. J. Lomas, and S. J. Rees. A simple model of PV system performance and its use in fault detection. *Solar Energy*, 84(4):624–635, April 2010. ISSN 0038-092X. , 10.1016/j.solener.2009.08.004. URL https://www.sciencedirect.com/science/article/pii/S0038092X0900187X.

[193] Radu Platon, Jacques Martel, Norris Woodruff, and Tak Y. Chau. Online Fault Detection in PV Systems. *IEEE Transactions on Sustainable Energy*, 6(4):1200–1207, October 2015. ISSN 1949-3037. , 10.1109/TSTE.2015.2421447.

[194] Mahmoud Dhimish and Violeta Holmes. Fault detection algorithm for grid-connected photovoltaic plants. *Solar Energy*, 137:236–245, November 2016. ISSN 0038-092X. , 10.1016/j.solener.2016.08.021. URL https://www.sciencedirect.com/science/article/pii/S0038092X16303486.

[195] A. Drews, A. C. de Keizer, H. G. Beyer, E. Lorenz, J. Betcke, W. G. J. H. M. van Sark, W. Heydenreich, E. Wiemken, S. Stettler, P. Toggweiler, S. Bofinger, M. Schneider, G. Heilscher, and D. Heinemann. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. *Solar Energy*, 81(4):548–564, April 2007. ISSN 0038-092X. , 10.1016/j.solener.2006.06.019. URL https://www.sciencedirect.com/science/article/pii/S0038092X06002040.

[196] Vesna Dimitrievska, Federico Pittino, Wolfgang Muehleisen, Nicole Diewald, Markus Hilweg, Andràs Montvay, and Christina Hirschl. Statistical Methods for Degradation Estimation and Anomaly Detection in Photovoltaic Plants. *Sensors*, 21(11):3733, January 2021. , 10.3390/s21113733. URL https://www.mdpi.com/1424-8220/21/11/3733.

[197] Rosemary Tawn, Jethro Browell, and Iain Dinwoodie. Missing Data in Wind Farm Time Series: Properties and Effect on Forecasts. *Electric Power Systems Research*, page 8, 2020. , 10.1016/j.epsr.2020.106640.

[198] Andreas Livera, Alexander Phinikarides, George Makrides, and George E. Georghiou. Impact of Missing Data on the Estimation of Photovoltaic System Degradation Rate. In *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, pages 1954–1958, June 2017. , 10.1109/PVSC.2017.8366442.

[199] Tahasin Shireen, Chenhui Shao, Hui Wang, Jingjing Li, Xi Zhang, and Mingyang Li. Iterative multi-task learning for time-series modeling of solar panel PV outputs. *Applied Energy*, 212:654–662, February 2018. ISSN 0306-2619. , 10.1016/j.apenergy.2017.12.058. URL https://www.sciencedirect.com/science/article/pii/S0306261917317737.

[200] Bella Espinar, Philippe Blanc, Lucien Wald, Carsten Hoyer-Klick, Marion Schroedter Homscheidt, and Thomas Wanderer. On quality control procedures for solar radiation and meteorological measures, from subhourly to montly average time periods. In *EGU General Assembly 2012*, April 2012. URL https://hal-mines-paristech.archives-ouvertes.fr/hal-00691350/document.

[201] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982. ISSN 0018-9448. , 10.1109/TIT.1982.1056489. URL http://ieeexplore.ieee.org/document/1056489/.

[202] Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD-96*, page 6, 1996.

[203] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*,

Lecture Notes in Computer Science, pages 160–172, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-37456-2. , 10.1007/978-3-642-37456-2_14.

[204] N. A. Engerer and F. P. Mills. KPV: A clear-sky index for photovoltaics. *Solar Energy*, 105: 679–693, July 2014. ISSN 0038-092X. , 10.1016/j.solener.2014.04.019. URL https://www.sciencedirect.com/science/article/pii/S0038092X14002151.

[205] Xixi Sun, Jamie M. Bright, Christian A. Gueymard, Brendan Acord, Peng Wang, and Nicholas A. Engerer. Worldwide performance assessment of 75 global clear-sky irradiance models using Principal Component Analysis. *Renewable and Sustainable Energy Reviews*, 111:550–570, September 2019. ISSN 13640321. , 10.1016/j.rser.2019.04.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1364032119302187.

[206] R Bird and Roland Hulstrom. A Simplified Clear Sky Model for Direct and Diffuse Insolation on Horizontal Surfaces. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1981.

[207] Christelle Rigollier, Olivier Bauer, and Lucien Wald. On the clear sky model of the ESRA — European Solar Radiation Atlas — with respect to the heliosat method. *Solar Energy*, 68 (1):33–48, January 2000. ISSN 0038-092X. , 10.1016/S0038-092X(99)00055-9. URL http://www.sciencedirect.com/science/article/pii/S0038092X99000559.

[208] Peder Bacher, Henrik Madsen, Bengt Perers, and Henrik Aalborg Nielsen. A non-parametric method for correction of global radiation observations. *Solar Energy*, 88:13–22, February 2013. ISSN 0038-092X. , 10.1016/j.solener.2012.10.024. URL http://www.sciencedirect.com/science/article/pii/S0038092X12003891.

[209] Jan Remund, Lucien Wald, Mireille Lefèvre, Thierry Ranchin, and John Page. Worldwide Linke turbidity information. *ISES Solar World Congress*, page 14, 2003.

[210] John A. Davies and Donald C. McKay. Estimating solar irradiance and components. *Solar Energy*, 29(1):55–64, January 1982. ISSN 0038-092X. , 10.1016/0038-092X(82)90280-8. URL https://www.sciencedirect.com/science/article/pii/0038092X82902808.

[211] Christian A. Gueymard. REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset. *Solar Energy*, 82(3):272–285, March 2008. ISSN 0038092X. , 10.1016/j.solener.2007.04.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X07000990.

[212] Mireille Lefèvre, Armel Oumbe, Philippe Blanc, Bella Espinar, Benoît Gschwind, Zhipeng Qu, Lucien Wald, Marion Schroedter Homscheidt, Carsten Hoyer-Klick, Antti Arola, Angela Benedetti, Johannes W. Kaiser, and Jean-Jacques Morcrette. McClear: A new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques*, 6:2403–2418, 2013. , 10.5194/amt-6-2403-2013.

[213] Dazhi Yang. Choice of clear-sky model in solar forecasting. *Journal of Renewable and Sustainable Energy*, 12(2):026101, March 2020. ISSN 1941-7012. , 10.1063/5.0003495. URL http://aip.scitation.org/doi/10.1063/5.0003495.

[214] Pierre Ineichen and Richard Perez. A new airmass independent formulation for the Linke turbidity coefficient. *Solar Energy*, 73(3):151–157, September 2002. ISSN 0038-092X. , 10.1016/S0038-092X(02)00045-2. URL https://www.sciencedirect.com/science/article/pii/S0038092X02000452.

[215] X. M. Chen, Y. Li, and R. Z. Wang. Performance study of affine transformation and the advanced clear-sky model to improve intra-day solar forecasts. *Journal of Renewable and*

*Sustainable Energy*, 12(4):043703, July 2020. ISSN 1941-7012. , 10.1063/5.0009155. URL http://aip.scitation.org/doi/10.1063/5.0009155.

[216] Christian A. Gueymard. Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment. *Solar Energy*, 74(5): 355–379, May 2003. ISSN 0038-092X. , 10.1016/S0038-092X(03)00195-6. URL https://www.sciencedirect.com/science/article/pii/S0038092X03001956.

[217] Joakim Munkhammar and Joakim Widén. Copula correlation modeling of aggregate solar irradiance in spatial networks. November 2016.

[218] Christopher J. Smith, Jamie M. Bright, and Rolf Crook. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. *Solar Energy*, 144:10–21, March 2017. ISSN 0038-092X. , 10.1016/j.solener.2016.12.055. URL https://www.sciencedirect.com/science/article/pii/S0038092X16306624.

[219] Clifford Hansen. Validation of Simulated Irradiance and Power for the Western Wind and Solar Integration Study, Phase II. Technical report, Sandia National Laboratories, October 2012.

[220] Gabi Friesen, Werner Herrmann, Giorgio Belluardo, and Bert Herteleer. *Photovoltaic module energy yield measurements: existing approaches and best practice: International Energy Agency Photovoltaic Power Systems Programme: IEA PVPS Task 13*. International Energy Agency IEA, Paris, 2018. ISBN 978-3-906042-52-7.

[221] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, March 2013. ISSN 1384-5810, 1573-756X. , 10.1007/s10618-012-0250-5. URL http://link.springer.com/10.1007/s10618-012-0250-5.

[222] D. Horvatic, H. E. Stanley, and B. Podobnik. Detrended cross-correlation analysis for non-stationary time series with periodic trends. *EPL (Europhysics Letters)*, 94(1):18007, April 2011. ISSN 0295-5075, 1286-4854. , 10.1209/0295-5075/94/18007. URL https://iopscience.iop.org/article/10.1209/0295-5075/94/18007.

[223] Cyril Voyant, Jan G. De Gooijer, and Gilles Notton. Periodic autoregressive forecasting of global solar irradiation without knowledge-based model implementation. *Solar Energy*, 174: 121–129, November 2018. ISSN 0038-092X. , 10.1016/j.solener.2018.08.076. URL https://www.sciencedirect.com/science/article/pii/S0038092X18308466.

[224] Robert Cleveland, William Cleveland, Jean McRae, and Irma Terpenning. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6(1):3–73, 1990. URL https://www.wessa.net/download/stl.pdf.

[225] Dazhi Yang, Panida Jirutitijaroen, and Wilfred M. Walsh. Hourly solar irradiance time series forecasting using cloud cover index. *Solar Energy*, 86(12):3531–3543, December 2012. ISSN 0038-092X. , 10.1016/j.solener.2012.07.029. URL http://www.sciencedirect.com/science/article/pii/S0038092X12003039.

[226] D. AlHakeem, P. Mandal, A. U. Haque, A. Yona, T. Senjyu, and T. Tseng. A new strategy to quantify uncertainties of wavelet-GRNN-PSO based solar PV power forecasts using bootstrap confidence intervals. In *2015 IEEE Power Energy Society General Meeting*, pages 1–5, July 2015. , 10.1109/PESGM.2015.7286233.

[227] Yuxin Wen, Donna AlHakeem, Paras Mandal, Shantanu Chakraborty, Yuan-Kang Wu, Tomonobu Senjyu, Sumit Paudyal, and Tzu-Liang Tseng. Performance Evaluation of Prob-

abilistic Methods Based on Bootstrap and Quantile Regression to Quantify PV Power Point Forecast Uncertainty. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4): 1134–1144, April 2020. ISSN 2162-2388. , 10.1109/TNNLS.2019.2918795.

[228] Juan Ospina, Alvi Newaz, and M. Omar Faruque. Forecasting of PV plant output using hybrid wavelet-based LSTM-DNN structure model. *IET Renewable Power Generation*, 13(7): 1087–1095, 2019. ISSN 1752-1424. , https://doi.org/10.1049/iet-rpg.2018.5779. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2018.5779.

[229] Manel Rhif, Ali Ben Abbes, Imed Riadh Farah, Beatriz Martínez, and Yanfang Sang. Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review. *Applied Sciences*, 9(7):1345, January 2019. , 10.3390/app9071345. URL https://www.mdpi.com/2076-3417/9/7/1345.

[230] Philippe Lauret, Cyril Voyant, Ted Soubdhan, Mathieu David, and Philippe Poggi. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, 112:446–457, February 2015. ISSN 0038-092X. , 10.1016/j.solener.2014.12.014. URL http://www.sciencedirect.com/science/article/pii/S0038092X14006057.

[231] Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, October 1992. ISSN 0304-4076. , 10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y.

[232] Wayne A. Fuller. *Introduction to Statistical Time Series*. John Wiley & Sons, September 2009. ISBN 978-0-470-31775-4.

[233] François Roueff and Andres Sanchez-Perez. Prediction of weakly locally stationary processes by auto-regression. *arXiv:1602.01942 [math, stat]*, January 2018. URL http://arxiv.org/abs/1602.01942.

[234] Henrik Madsen. *Time Series Analysis*. Chapman & Hall, 2007. ISBN 978-0-429-19583-9.

[235] Richard Perez, Sergey Kivalov, Jim Schlemmer, Karl Hemker, and Thomas E. Hoff. Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance. *Solar Energy*, 86(8):2170–2176, August 2012. ISSN 0038092X. , 10.1016/j.solener.2012.02.027. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X12000928.

[236] K. Bellinguer, R. Girard, G. Bontron, and G. Kariniotakis. Short-Term Photovoltaic Generation Forecasting Using Multiple Heterogenous Sources of Data. In *36th European Photovoltaic Solar Energy Conference and Exhibition*, pages 1422–1427, October 2019. , 10.4229/EUPVSEC20192019-5DO.2.4. URL http://www.eupvsec-proceedings.com/proceedings?paper=48534.

[237] Richard Perez, Mathieu David, Thomas E. Hoff, Mohammad Jamaly, Sergey Kivalov, Jan Kleissl, Philippe Lauret, and Marc Perez. Spatial and Temporal Variability of Solar Energy. *Foundations and Trends® in Renewable Energy*, 2016. , 10.1561/2700000006. URL https://hal.archives-ouvertes.fr/hal-01467044.

[238] Christopher K. Wikle, Andrew Zammit-Mangion, and Noel A. C. Cressie. *Spatio-temporal statistics with R*. Chapman & Hall/CRC the R series. CRC Press, Taylor & Francis Group, Boca Raton, 2019. ISBN 978-1-138-71113-6.

[239] Akın Taşcıkaraoğlu. Impacts of Accurate Renewable Power Forecasting on Optimum Operation of Power System. In *Optimization in Renewable Energy Systems*, pages 159–175. Elsevier,

2017. ISBN 978-0-08-101041-9. URL https://linkinghub.elsevier.com/retrieve/pii/B9780081010419000053.

[240] Songjian Chai, Zhao Xu, Youwei Jia, and Wai Kin Wong. A Robust Spatiotemporal Forecasting Framework for Photovoltaic Generation. *IEEE Transactions on Smart Grid*, pages 1–1, 2020. ISSN 1949-3061. , 10.1109/TSG.2020.3006085.

[241] M. Heidari Kapourchali, M. Sepehry, and V. Aravinthan. Multivariate Spatio-temporal Solar Generation Forecasting: A Unified Approach to Deal With Communication Failure and Invisible Sites. *IEEE Systems Journal*, pages 1–9, 2018. ISSN 1932-8184. , 10.1109/JSYST.2018.2869825.

[242] Aloysius W. Aryaputera, Dazhi Yang, Lu Zhao, and Wilfred M. Walsh. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. *Solar Energy*, 122:1266–1278, December 2015. ISSN 0038092X. , 10.1016/j.solener.2015.10.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X15005745.

[243] Mohammad Jamaly and Jan Kleissl. Spatiotemporal interpolation and forecast of irradiance data using Kriging. *Solar Energy*, 158:407–423, December 2017. ISSN 0038092X. , 10.1016/j.solener.2017.09.057. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X17308447.

[244] Dazhi Yang. Ultra-fast preselection in lasso-type spatio-temporal solar forecasting problems. *Solar Energy*, 176:788–796, December 2018. ISSN 0038-092X. , 10.1016/j.solener.2018.08.041. URL http://www.sciencedirect.com/science/article/pii/S0038092X18308120.

[245] Chao Huang, Long Wang, and Loi Lei Lai. Data-Driven Short-Term Solar Irradiance Forecasting Based on Information of Neighboring Sites. *IEEE Transactions on Industrial Electronics*, 66(12):9918–9927, December 2019. ISSN 0278-0046, 1557-9948. , 10.1109/TIE.2018.2856199. URL https://ieeexplore.ieee.org/document/8415759/.

[246] Caroline Persson, Peder Bacher, Takahiro Shiga, and Henrik Madsen. Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, 150:423–436, July 2017. ISSN 0038092X. , 10.1016/j.solener.2017.04.066. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X17303717.

[247] S. Cros, O. Liandrat, N. Sébastien, and N. Schmutz. Extracting cloud motion vectors from satellite images for solar power forecasting. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 4123–4126, July 2014. , 10.1109/IGARSS.2014.6947394.

[248] L. Mazorra Aguiar, B. Pereira, M. David, F. Díaz, and P. Lauret. Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks. *Solar Energy*, 122:1309–1324, December 2015. ISSN 0038-092X. , 10.1016/j.solener.2015.10.041. URL http://www.sciencedirect.com/science/article/pii/S0038092X15005927.

[249] David P. Larson and Carlos F. M. Coimbra. Direct Power Output Forecasts From Remote Sensing Image Processing. *Journal of Solar Energy Engineering*, 140(021011), February 2018. ISSN 0199-6231. , 10.1115/1.4038983. URL https://doi.org/10.1115/1.4038983.

[250] Z. Si, Y. Yu, M. Yang, and P. Li. Hybrid Solar Forecasting Method Using Satellite Visible Images and Modified Convolutional Neural Networks. *IEEE Transactions on Industry Applications*, 57(1):5–16, January 2021. ISSN 1939-9367. , 10.1109/TIA.2020.3028558.

[251] Hanchuan Peng, Fuhui Long, and C. Ding. Feature Selection based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005. ISSN 0162-8828. , 10.1109/TPAMI.2005.159. URL http://ieeexplore.ieee.org/document/1453511/.

[252] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, August 1987. ISSN 0169-7439. , 10.1016/0169-7439(87)80084-9. URL https://www.sciencedirect.com/science/article/pii/0169743987800849.

[253] Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, 2021. ISSN 0169-2070. , 10.1016/j.ijforecast.2020.07.007. URL https://www.sciencedirect.com/science/article/pii/S0169207020301114.

[254] Yann LeCun, Yoshua Bengio, and T Bell Laboratories. Convolutional Networks for Images, Speech, and Time-Series. In *The handbook of brain theory and neural networks*, page 14. MIT Press, m.a. arbib edition, 1995.

[255] Happy Aprillia, Hong-Tzer Yang, and Chao-Ming Huang. Short-Term Photovoltaic Power Forecasting Using a Convolutional Neural Network–Salp Swarm Algorithm. *Energies*, 13(8): 1879, January 2020. , 10.3390/en13081879. URL https://www.mdpi.com/1996-1073/13/8/1879.

[256] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *arXiv:1506.04214 [cs]*, June 2015. URL http://arxiv.org/abs/1506.04214.

[257] Sujan Ghimire, Ravinesh C. Deo, Nawin Raj, and Jianchun Mi. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied Energy*, 253:113541, November 2019. ISSN 0306-2619. , 10.1016/j.apenergy.2019.113541. URL http://www.sciencedirect.com/science/article/pii/S0306261919312152.

[258] Vishnu Suresh, Przemyslaw Janik, Jacek Rezmer, and Zbigniew Leonowicz. Forecasting Solar PV Output Using Convolutional Neural Networks with a Sliding Window Algorithm. *Energies*, 13(3):723, January 2020. , 10.3390/en13030723. URL https://www.mdpi.com/1996-1073/13/3/723.

[259] Ali Agga, Ahmed Abbou, Moussa Labbadi, and Yassine El Houm. Short-term self consumption PV plant power production forecasts based on hybrid CNN-LSTM, ConvLSTM models. *Renewable Energy*, 177:101–112, November 2021. ISSN 0960-1481. , 10.1016/j.renene.2021.05.095. URL https://www.sciencedirect.com/science/article/pii/S0960148121007771.

[260] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Li Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent Advances in Convolutional Neural Networks. *arXiv:1512.07108 [cs]*, October 2017. URL http://arxiv.org/abs/1512.07108.

[261] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980.

[262] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016. ISSN 2196-1115. , 10.1186/s40537-016-0043-6. URL https://doi.org/10.1186/s40537-016-0043-6.

[263] Alfredo Nespoli, Emanuele Ogliari, Sonia Leva, Alessandro Massi Pavan, Adel Mellit, Vanni Lughi, and Alberto Dolara. Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques. *Energies*, 12(9):1621, April 2019. ISSN 1996-1073. , 10.3390/en12091621. URL https://www.mdpi.com/1996-1073/12/9/1621.

[264] Claudio Monteiro, Tiago Santos, L. Alfredo Fernandez-Jimenez, Ignacio J. Ramirez-Rosado, and M. Sonia Terreros-Olarte. Short-Term Power Forecasting Model for Photovoltaic Plants

Based on Historical Similarity. *Energies*, 6(5):2624–2643, May 2013. , 10.3390/en6052624. URL https://www.mdpi.com/1996-1073/6/5/2624.

[265] Patrick Mathiesen, John M. Brown, and Jan Kleissl. Geostrophic Wind Dependent Probabilistic Irradiance Forecasts for Coastal California. *IEEE Transactions on Sustainable Energy*, 4(2):510–518, April 2013. ISSN 1949-3037. , 10.1109/TSTE.2012.2200704.

[266] Bastien Alonzo, Peter Tankov, Philippe Drobinski, and Riwal Plougonven. Probabilistic wind forecasting up to three months ahead using ensemble predictions for geopotential height. *International Journal of Forecasting*, 36(2):515–530, April 2020. ISSN 01692070. , 10.1016/j.ijforecast.2019.07.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207019302018.

[267] Pascal Horton, Michel Jaboyedoff, and Charles Obled. Using genetic algorithms to optimize the analogue method for precipitation prediction in the Swiss Alps. *Journal of Hydrology*, 556:1220–1231, January 2018. ISSN 0022-1694. , 10.1016/j.jhydrol.2017.04.017. URL https://www.sciencedirect.com/science/article/pii/S0022169417302391.

[268] B. O. Akyurek, A. S. Akyurek, J. Kleissl, and T. S Rosing. TESLA - Taylor expanded solar analog forecasting. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 127–132, 2014. , 10.1109/SmartGridComm.2014.7007634.

[269] Luca Delle Monache, F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight. Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, 141(10):3498–3516, May 2013. ISSN 0027-0644. , 10.1175/MWR-D-12-00281.1. URL https://journals.ametsoc.org/doi/10.1175/MWR-D-12-00281.1.

[270] William S. Cleveland and Clive Loader. Smoothing by Local Regression: Principles and Methods. In Wolfgang Härdle and Michael G. Schimek, editors, *Statistical Theory and Computational Aspects of Smoothing*, Contributions to Statistics, pages 10–49, Heidelberg, 1996. Physica-Verlag HD. ISBN 978-3-642-48425-4.

[271] Charles Obled, Guillaume Bontron, and Rémy Garçon. Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, 63(3-4):303–324, 2002. ISSN 01698095. , 10.1016/S0169-8095(02)00038-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0169809502000388.

[272] R. Romero, G. Sumner, C. Ramis, and A. Genovés. A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area. *International Journal of Climatology*, 19(7):765–785, 1999. ISSN 1097-0088. , 10.1002/(SICI)1097-0088(19990615)19:7<765::AID-JOC388>3.0.CO;2-T. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0088%2819990615%2919%3A7%3C765%3A%3AAID-JOC388%3E3.0.CO%3B2-T.

[273] M. H. Ramos, S. J. van Andel, and F. Pappenberger. Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, 17(6):2219–2232, June 2013. ISSN 1607-7938. , 10.5194/hess-17-2219-2013. URL https://hess.copernicus.org/articles/17/2219/2013/.

[274] Hilaf Hasson and Yuyang Wang. Probabilistic Forecasting: A Level-Set Approach. In *NeurIPS 2021*, page 13, 2021.

[275] P. Pinson and G. Kariniotakis. On-line assessment of prediction risk for wind power production forecasts. *Wind Energy*, 7(2):119–132, 2004. ISSN 1099-1824. , https://doi.org/10.1002/we.114. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.114.

[276] Roger Koenker and Gilbert Bassett. Regression Quantiles. *Econometrica*, 46(1):33, January 1978. ISSN 00129682. , 10.2307/1913643. URL https://www.jstor.org/stable/1913643?origin=crossref.

[277] Nicolai Meinshausen. Quantile Regression Forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006. ISSN ISSN 1533-7928. URL http://jmlr.org/papers/v7/meinshausen06a.html.

[278] Mathieu David, Josselin Le Gal La Salle, Faly H. Ramahatana Andriamasomanana, and Philippe Lauret. Probabilistic Solar Forecasts Evaluation Part 1: Ensemble Prediction Systems (EPS). In *Proceedings of the ISES Solar World Congress 2019*, pages 1–9, Santiago, Chile, 2019. International Solar Energy Society. ISBN 978-3-9820408-1-3. , 10.18086/swc.2019.43.02. URL http://proceedings.ises.org/citation?doi=swc.2019.43.02.

[279] Pierre Pinson, Henrik Aa Nielsen, Jan K. Møller, Henrik Madsen, and George N. Kariniotakis. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*, 10(6):497–516, 2007. ISSN 1099-1824. , 10.1002/we.230. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.230.

[280] Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, May 2005. ISSN 1520-0493, 0027-0644. , 10.1175/MWR2904.1. URL https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2904.1.xml.

[281] James E. Matheson and Robert L. Winkler. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096, June 1976. ISSN 0025-1909, 1526-5501. , 10.1287/mnsc.22.10.1087. URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.22.10.1087.

[282] Hans Hersbach. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570, October 2000. ISSN 0882-8156, 1520-0434. , 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

[283] Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. , 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

[284] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[285] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 2014. URL http://arxiv.org/abs/1406.2661.

[286] Kai Qu, Gangquan Si, Zihan Shan, XiangGuang Kong, and Xin Yang. Short-term forecasting for multiple wind farms based on transformer model. *Energy Reports*, 8:483–490, August 2022. ISSN 2352-4847. , 10.1016/j.egyr.2022.02.184. URL https://www.sciencedirect.com/science/article/pii/S2352484722004310.

[287] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press book, mit press edition, 2016. URL http://www.deeplearningbook.org/.

[288] Francois Chollet. Keras, 2015. URL https://github.com/fchollet/keras.

[289] Leonard Kaufman and Peter J. Rousseeuw. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*, pages 68–125. John Wiley & Sons, Ltd, 1990. ISBN 978-0-470-31680-1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2.

[290] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. ISSN 0035-9254. , 10.2307/2346830. URL https://www.jstor.org/stable/2346830.

## RÉSUMÉ

Dans un contexte d'épuisement des ressources naturelles, les sources d'énergies renouvelables jouent un rôle croissant dans le mix de la production électrique. Cependant, une part importante des renouvelables peut compromettre la stabilité du réseau électrique en raison de leurs variabilités. Il est donc primordial de connaître la quantité d'énergie future produite afin d'assurer l'équilibre entre production et consommation. Cette thèse porte sur l'amélioration de la précision des prévisions court-terme de la production photovoltaïque. Pour y parvenir, un couplage entre modèles statistiques et modèles physiques est proposé, en plus d'une architecture permettant de conditionner les modèles à la situation météorologique. En outre, un large éventail de sources d'information est considéré. A cet égard, une analyse approfondie des données permet de mettre en exergue l'information pertinente ainsi que les dépendances spatio-temporelles pouvant exister entre les différentes variables.

## MOTS CLÉS

Prévision court-terme de la production photovoltaïque, Données spatio-temporelle, Conditionement par la situation météorologique, Observations infrarouges, Correction de données aberrantes, Systèmes électriques intelligents

## ABSTRACT

In a context of natural resources depletion, weather-dependent renewable energy sources play an increasingly important role in the electricity generation mix. Yet, high shares of renewables can jeopardise the safe operation of the power grid due to their variable nature. To address this challenge, it is essential to know the future amount of energy produced to balance production and consumption. In this thesis, we explore two main approaches that aim at improving the accuracy of short-term photovoltaic generation forecasting. The first option is to extend the existing statistical models found in the literature through the coupling with a physics-based model, and by operating a shift from static to weather-adaptive models. The second option lies in extending the range of available sources of information. In this regard, an in-depth quality analysis of production measurements emphasises relevant information, and exhibits the spatio-temporal correlations that may exist between the inputs.

## KEYWORDS

Short-term photovoltaic generation forecasting, Spatio-temporal data, Weather conditioning, Infrared-based observations, Correction of fallacious observations, Smart grids