Study and enhancement of Bayesian calibration applied to building energy models

Samih Akkari

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Mines Paris-PSL

Study and enhancement of Bayesian calibration applied to building energy models

Étude et amélioration de l'application de calibrage bayésien dans les modèles énergétiques du bâtiment

Soutenue par

Samih Akkari

Le 03 Octobre 2022

Ecole doctorale n° 621

Ingénierie des Systèmes, Matériaux, Mécanique et Énergétique

Spécialité

Énergétique et Génie des Procédés

Composition du jury :

Laurent MORA Professeur, Université de Bordeaux	*Président*
Ruchi Choudhary Professeur, Université de Cambridge	*Rapporteur*
Stéphane PLOIX Professeur, Université de Grenoble	*Rapporteur*
Séverine DEMEYER Docteur, LNE	*Examinateur*
Christian ROBERT Professeur, Université Paris Dauphine	*Examinateur*
Maxime TROCME Docteur, VINCI	*Examinateur*
Bruno PEUPORTIER Directeur de recherche, Mines Paris	*Directeur de thèse*
Patrick SCHALBART Docteur, Mines Paris	*Examinateur*

MINES PARIS | PSL

Table of content

Nomenclature

List of symbols

- A_0, A_k, B_k: Fourier analysis coefficients [-]
- A_1: Upper asymptote [-]
- $\hat{A}_1$: Reduced upper asymptote [-]
- A_2: Lower asymptote [-]
- a: Slope [-]
- B_l: Left sampling bound [-]
- B_r: Right sampling bound [-]
- c: Intercept [-]
- d^*: Euclidean distance [-]
- d_{dist}: Distance between posteriors and true values [-]
- D: Total variance of the model response [-]
- D_i: Model response variance caused by parameter θ_i [-]
- $D(\sim i)$: Model response variance caused by all parameter except θ_i [-]
- e_i: Exposure coefficient [-]
- f: Frequency [-]
- f_{max}: Maximum frequency [-]
- $f(\theta)$: Model response at parameter θ [°C]
- $f_t(\cdot)$: Intermediate distributions in SMC samplers [-]
- G: Transformation function [-]
- $I(.)$: Criterion applied to each node [-]
- k: Search parameter [-]
- $K(\cdot)$: Kernel function [-]
- $l_b(Y)$: Leaf in which the sample *y* landed [-]
- M: Harmonics (FAST sensitivity analysis) [-]
- N: Number of samples [-]
- N_{steps}: Number of MCMC jumps in one iteration [-]

- N_α: Particles with distances less than α-quantile [-]
- $N(\cdot)$: Normal distribution [-]
- N_{min}: Minimum number of samples in a node [-]
- N^1: Sample size to train the first RF in ARF [-]
- N^t: Sample size at iteration t [-]
- n_b: Number of samples in leaf $l_b(Y)$ [-]
- n_c: Number of concordant pairs [-]
- n_d: Number of discordant pairs [-]
- n_{50}: Air change rate per hour [vol/hr]
- n_{m_r}: Number of samples in right node [-]
- n_{m_l}: Number of samples in left node [-]
- n_{try}: Subset on which the criteria is minimised [-]
- n^t: Number of newly generated samples at each iteration in ARF [-]
- $p(\theta)$: Prior distribution of parameter θ [-]
- $p(\theta|Z)$: Likelihood; distribution of parameter θ given data Z [-]
- $p(Z)$: Normalising factor in Bayes law [-]
- p: Covariance scaling parameter [-]
- $P_{acc_{min}}$: Convergence threshold [-]
- Q_{infilt}: Infiltration flow rate [m^3/hr]
- r_{θ_i}: Ranking of parameter θ_i by Sobol method [-]
- x_{θ_i}: Ranking of parameter θ_i by either Morris or RBD-FAST [-]
- R_L: Residual matrix [-]
- S: Sensitivity matrix [-]
- s: Shift from true values [-]
- $S(\cdot)$: Summary statistic [-]
- S_i: First-order sensitivity index [-]
- S_{ij}: Second-order sensitivity index [-]
- S^c: Corrected sensitivity index [-]
- S_m: Mean sensitivity index [-]
- S_{std}: Standard deviation sensitivity index [-]

- S_L: Sensitivity matrix comprising L columns [-]
- S_P: Projected sensitivity matrix [-]
- TS: Total sensitivity index [-]
- T: Temperature [℃]
- t: Time [hr]
- $U(\cdot)$: Uniform distribution [-]
- V: Volume of the heated space [m^3]
- $V(Y)$: Weighted variance of the random forest posterior [-]
- w: Weights [-]
- $\bar{y}_m$: The average value of the responses in node m [-]
- y_i: Response of each sample in random forest data set [-]

List of greek letters

- α: Acceptance rejection ratio [-]
- β_t: Annealing parameter at iteration t [-]
- δ: Tolerance [℃]
- $\delta(x)$: Model discrepancy [-]
- Δt: Difference between two time steps [hr]
- $\in_i$: Wind speed correction coefficient [-]
- $\varepsilon(x)$: Observation errors [-]
- η: Model [-]
- θ: Parameter [-]
- θ^*: Perturbed or proposal sample [-]
- $\bar{\theta}_t$: Weighted samples mean of parameter θ at iteration t [-]
- μ: Average over all trees predictions [-]
- μ^*: Absolute mean of elementary effects [℃]
- μ_b: Prediction of tree b [-]
- μ_θ: Prior mean of parameter θ [-]
- $\rho(x, y)$: Distance between measurements and predictions [℃]

- σ_j: Sample variance of parameter j [-]
- σ_θ: Prior standard deviation of parameter θ [-]
- Σ_t: Samples covariance at iteration t [-]
- τ: Kandall tau correlation [-]

Abbreviations

- ABC: Approximate Bayesian computation [-]
- ABC-RF: Random forest for approximate Bayesian computation [-]
- APMC: Adaptive population Monte Carlo [-]
- AIC: Akaike information criterion [-]
- ARF: Adaptive random forest [-]
- DAISY: Differential algebra for identifiability of systems [-]
- DBEM: Dynamic building energy model [-]
- CMV: Controlled mechanical ventilation [-]
- CATMIP: Cascading adaptive transitional metropolis in parallel [-]
- DIC: Deviance information criterion [-]
- EPC: Energy performance contracting [-]
- ESCOs: Energy sevice companies [-]
- ESS: Effective sample size [-]
- FAST: Fourier amplitude sensitivity test [-]
- EFAST: Extended FAST [-]
- HVAC: Heating ventilation and air conditioning [-]
- ID: Identifiability distance indicator [-]
- JS: Janson-Shannon [-]
- KL: Kullback-Leibler [-]
- MCMC: Markov chain Monte Carlo [-]
- TMCMC : Transitional MCMC [-]
- PCA: Principle component analysis [-]
- PCC: Pearson correlation coefficient [-]
- PMC: Population Monte Carlo [-]

- RBD: Random balance design [-]
- RMSE: Root-mean-square-error [-]
- RF: Random forest [-]
- SMC: Sequential Monte Carlo [-]
- SRC: Standardised regression coefficient [-]
- SRRC: Standardised rank regression coefficient [-]
- TMCMC: Transitional Marchov chain Monte Carlo [-]

General introduction

The world has experienced a significant increase in the energy consumption in the recent years (Marchi and Zanoni 2017). Energy efficiency measures have been employed in different sectors such as the industrial sector since it plays a big role in the energy consumption. The building sector is responsible for 40 % of the total energy consumption and 36 % of the greenhouse gases emissions in the EU[1]. As stated by the European commission[1], roughly 75 % of the EU building stock is energy inefficient which explains its huge role in energy consumption. Recently, efforts have been focused on developing new measures and strategies in the building sector (Ruggeri et al., 2020). The rate at which buildings are renovated in the EU is less than 1 % of the national building stock every year[1]. Sandberg et al. (2016) used a probabilistic dynamic building stock model to simulate the development of dwellings in different EU countries. They found that the renovation rate is very unlikely to grow higher than the current rates. Tuominen et al. (2012) summarised the main barriers against renovation reported by stakeholder interviews in different countries. Some reported barriers were related to financing. People's low income in certain countries prevented the application of such improvements. Another reason related to financing is that the pollution caused by existing building stocks is not included in the energy price; this makes people less prone to save energy. A common barrier is that people are ill-informed about energy efficiency and regulations which makes it a low priority for them in addition to the risk and inconvenience that they might feel about these works. Some reported barriers clearly show a lack of knowledge about the relation between energy efficiency and price and how cost effective renovation could be. In October 2020, the European Commission presented the strategy which aims at increasing the rate of building renovation[1].

Energy performance contracting (EPC) is a well established type of contract in which the customers do not bare any performance and technological risks. EPC is the contracting of a specialised energy service company (ESCO) to guarantee energy saving during the contract time. An ESCO performs a comprehensive energy audit of the concerned building and identifies energy efficient improvements that meet the customers' needs. It guarantees that the new measures will generate sufficient energy cost savings to finance the project before the end of

[1] https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-lut-17_en

the contract. After the contract ends, the customer will benefit from all the cost savings. A diagram of the EPC concept is presented in Figure 1.

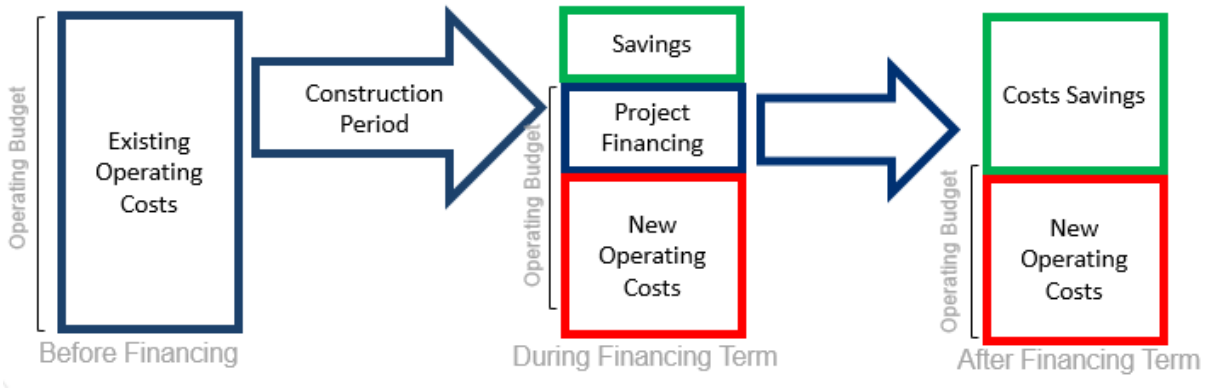

Figure 1: EPC concept[2]

Dynamic building energy models (DBEMs) are the main tools for ESCOs to simulate and analyse different possible improvements. Over the past decade, DBEMs have been increasingly used to estimate the energy behaviour of buildings and to predict the performance of newly designed buildings that are to be constructed, or existing buildings to be renovated. It can also be used to aid in designing appropriate HVAC systems to meet the needs of the buildings by the ability to test different control strategies. Control of building usages can easily be assessed using these tools in order to evaluate the building's performance subject to different conditions and operation (offices, residential buildings, etc.). These tools can be used for many more purposes as they are less expensive and much faster to run compared to an experimental campaign.

However, the results of these models are characterised by some degree of uncertainty and could show poor fit to the measured observations. Normally, this is the case in all numerical models as they share the same aspects and requirements that are themselves uncertain. Several factors serve as the source of uncertainty and inaccuracy in the numerical models predictions. These models and especially DBEMs are physics-based simulation tools built on simplified models of physical phenomena that are in reality more complicated. These simplifications and assumptions cause what is called model inadequacy, which is the discrepancy between the observations and simulations even if the case study with all its specifications is accurately defined. In practice, it is difficult to define precisely these specifications, which are also called the parameters of the DBEM. Thus, the accuracy and the confidence in the model predictions

[2] adapted from https://deq.mt.gov/energy/Programs/epc accessed 19/07/2022

is affected. In reality, there is no such detailed and accurate information about the building geometry, construction materials thermophysical properties, and its mechanical systems. Thus, the modeller will have to identify them with a degree of uncertainty and try to guess the best possible scenarios. Those uncertainties impacting the inputs will eventually propagate through the model and lower the confidence in its predictions, this is called parameter uncertainty.

These problems affect the level of confidence in guaranteeing the energy performance and thus puts a huge risk on the EPC provider. Due to this, uncertainty analysis has received an increasing attention in the field of building energy simulation (Tian et al. 2018). Uncertainty analysis can be classified into two types. The first type is the forward uncertainty analysis (uncertainty propagation) which aims at quantifying how much the uncertainty in the inputs contributes to the uncertainty in the outputs (i.e. how much certain one is about model's outputs). This differs from the sensitivity analysis in the manner that even if a specific parameter is very important in the model, it will not contribute much to the simulation output uncertainty if it is well known, since it will then be provided to the model as a constant value or as a very precise probability distribution. The second type is the inverse uncertainty analysis that is also called calibration. The aim is to diminish the uncertainties in the model predictions and to fit better to the actual behaviour of the building. It also estimates the most probable values of the uncertain parameters from collected in-situ measurements. This process provides the designer with a model calibrated on the same building as the one on which the retrofit will be applied which aids the performance guarantee process. It is important to mention that EPC is only one example where calibration is applied to enhance the confidence level.

At some point, manual calibration might be preferred against automatic calibration since the latter is a mathematical-based approach that could fail to attach physical reality to the uncertain parameters. However, it can be argued that automatic calibration performs better in fully exploring the parameter space, which makes it less prone to sub-optimal results compared to manual calibration. Bayesian calibration is an automated calibration method that combines both the in-situ measurements available and the prior knowledge about the building, to generate a model that fits better to data while accounting for the uncertainties in the predictions and model parameters. In other words, it naturally quantifies the uncertainties in the model predictions after calibration since the calibrated parameters are in the form of probability distributions.

This thesis is oriented towards enhancing the application of Bayesian calibration methods to building energy models. One of the main issues in calibration methods is that they could be computationally intensive. Accordingly, the building energy models are calibrated on a subset of the most influential parameters. A sensitivity analysis is applied to select these parameters. However, the sensitivity methods that provide a very precise ranking could also be computationally intensive. In the literature, the most used method is Morris since it provides a good approximation of the importance with a relatively low computational cost. RBD-FAST is another promising method in the field which is also computationally efficient. In chapter 2, a detailed comparison between Morris and RBD-FAST methods is conducted in terms of robustness, accuracy and computational efficiency using Sobol method as the reference method.

Bayesian methods can be divided into two main families: likelihood-dependent and likelihood-independent methods also called approximate Bayesian computation (ABC). In chapter 3, different methods are selected from the literature and applied to a virtual case study. The methods are then assessed in terms of accuracy and computational efficiency.

Another issue of calibration is the un-identifiability of the calibration parameters. Un-identifiability of a parameter means that the parameter cannot be identified from the data. The reason behind un-identifiability could be that the parameter itself is unimportant. It could also be that the set of parameters chosen for calibration have a significant degree of interactions, which means that there is no unique solution for the problem. It could also be related to the insufficiency or the poor quality of the data, which is not accounted for in this thesis. In chapter 1, the concept of identifiability analysis is explained in more details. In chapter 4, a sensitivity-based identifiability analysis is applied to select the parameters while accounting for their identifiability in addition to their importance.

The number of calibration parameters is also an important thing to account for. Too many parameters lead not only to more un-identifiability problems, but also make the calibration process computationally intensive. Too few parameters means that some influential ones might not be accounted for, affecting the precision of the calibration. However, selecting the appropriate number of parameters for a given case study is a hard task. In chapter 4 this issue is analysed by assessing the calibration performance with an increasing number of parameters.

In chapter 5, a new Bayesian method that belongs to the ABC family is proposed. This method applies the concepts from machine learning to find the best estimates of the calibration

parameters in form of distributions. It is applied to a virtual case study and compared to other existing methods in the literature. Finally, a real case study corresponding to real monitored data is used to evaluate the calibration methods.

Chapter 1

Context and state of art

The objective of this chapter is to highlight the main steps in the calibration methodology. Firstly, an overview on the sensitivity methods and its application on building energy models is conducted. Secondly, the concept of identifiability analysis is explained and an overview on different existing methods is provided. Finally, Bayesian calibration is presented and various sampling methods are introduced. This chapter serves as a background for all the subsequent chapter in this thesis.

Résumé du chapitre

Du fait que le secteur du bâtiment est responsable de la plus grande part de la consommation d'énergie en France et en Europe, les chercheurs s'intéressent de plus en plus à l'amélioration de l'efficacité énergétique des bâtiments. De nombreux pays ont établi des lignes directrices et des politiques à prendre en compte lors de la construction de nouveaux bâtiments pour s'assurer qu'ils sont conformes aux objectifs de performance énergétique. Cependant, le pourcentage de bâtiments nouvellement construits par rapport à ceux déjà existants est faible. Cela signifie qu'il y a un grand intérêt à rénover ces bâtiments existants. Les modèles énergétiques des bâtiments sont normalement utilisés pour faciliter la quantification et la comparaison de différentes mesures et de leurs économies d'énergie possibles. Une telle approche basée sur des modèles est soumise à de nombreuses sources d'incertitude. Cela signifie que ces modèles peuvent permettre des prises de décision pour des mesures de rénovation, cependant la gestion des risques ou l'estimation de la fiabilité nécessite des efforts supplémentaires.

Pour résoudre ce problème, le calibrage des modèles énergétiques des bâtiments améliore la précision de ces modèles dans la représentation du comportement réel du bâtiment étudié. D'autre part, le calibrage bayésien est une approche efficace pour quantifier les incertitudes dans les paramètres du modèle et les prédictions correspondantes du modèle sur les mesures de rénovation proposées, ce qui permet de représenter les économies d'énergie prévues dans le futur sous la forme d'une distribution de probabilité à partir de laquelle un niveau de confiance peut être évalué. En conséquence, ces dernières années, de nombreux chercheurs se sont concentrés sur ces approches de calibrage pour améliorer la performance des évaluations en termes de précision et d'efficacité de calcul. Dans cette thèse, une méthodologie de calibrage est étudiée sur l'ensemble de ses étapes.

Notre objectif principal dans cette thèse est de rendre compte d'une source d'incertitudes des modèles numériques qui est liée à la connaissance imprécise des paramètres qui ont une influence importante sur la simulation. Normalement, ces paramètres doivent décrire précisément le cas étudié pour pouvoir prédire avec précision la consommation d'énergie observée ou prévoir le comportement futur d'un bâtiment en cours de conception. Cependant, dans la pratique, la plupart de ces paramètres ne sont pas connus avec précision et ne peuvent être affectés d'une valeur spécifique en raison de diverses sources d'incertitude. Différentes études visaient à recueillir un maximum de données expérimentales sur les paramètres

habituellement utilisés dans la simulation du bâtiment. Par exemple, Clarke et al. (1991) ont fourni un examen complet des données disponibles dans le monde concernant les propriétés thermophysiques des matériaux utilisés dans la simulation en construction. Ces études constituent la base sur laquelle repose la quantification de l'incertitude des paramètres. La quantification de l'incertitude est une étape essentielle de la méthodologie d'analyse de l'incertitude et plus ces quantifications et distributions sont réalistes, plus nous pouvons être sûrs des limites d'incertitude de la sortie de simulation.

Normalement, dans le processus de calibrage, après avoir collecté et traité les données, un sous-ensemble de l'ensemble des paramètres est choisi pour être estimé à partir des mesures. Cela amène le besoin d'une analyse de sensibilité (AS) qui est souvent utilisée comme base pour l'analyse d'incertitude. Dans le contexte de la réalisation d'une analyse d'incertitude inverse, comme cela sera discuté dans les chapitres suivants, la demande de calcul devient coûteuse si le nombre de paramètres impliqués est trop grand. Ainsi, il est important d'analyser l'effet de tous les paramètres d'entrée du modèle sur les sorties du modèle pour effectuer l'analyse d'incertitude sur ceux qui sont les plus influents et d'écarter les paramètres dont les variations n'affectent pas la sortie du modèle. Dans ce contexte, rejeter signifie attribuer une valeur constante lors de l'analyse de l'incertitude.

L'analyse de sensibilité est une méthode statistique qui permet de simplifier ou de mieux comprendre les modèles numériques en classant et hiérarchisant les facteurs d'incertitude selon leur influence sur l'incertitude de sortie. Les indices de sensibilité sont calculés pour chaque facteur du modèle. Plus la valeur de l'indice de sensibilité est élevée, plus le paramètre est influent. Il existe un grand nombre de méthodes d'analyse de sensibilité et le choix de l'une ou de l'autre doit se faire en fonction des objectifs de l'étude. Pour une description détaillée des différentes méthodes et de leurs applications, le lecteur est renvoyé à Pannier et al. (2018).

En bref, la sensibilité locale est considérée comme l'une des méthodes de sensibilité les plus simples et est largement utilisée dans la littérature (Spitz et al. 2012). Cette méthode est une méthode un à la fois (OAT) puisque l'approche consiste à modifier un facteur à la fois et à effectuer la simulation pour détecter le changement dans la sortie de simulation causé uniquement par le facteur perturbé. Cette méthode est classée dans les méthodes qualitatives car elle ne donne qu'une information qualitative sur l'importance de chaque paramètre. Un autre ensemble de méthodes est constitué par les méthodes de dépistage. Une méthode de dépistage largement utilisée est la méthode de Morris (Morris, 1991). Elle est classée dans les méthodes

« une à la fois » où, pour chaque exécution du modèle, un seul paramètre d'entrée est modifié. Elle est plus précise que les méthodes locales car elle tient compte de la non-linéarité et des interactions entre les paramètres : la sensibilité d'un paramètre peut dépendre des valeurs d'autres paramètres.

Les méthodes de sensibilité globale évaluent quantitativement l'importance de chaque paramètre sur l'ensemble de l'espace d'entrée en tenant compte de l'interaction entre les paramètres. Ainsi, elles explorent tout l'espace des paramètres mais peuvent être considérablement intensives en calcul contrairement aux méthodes locales et de dépistage. La raison en est qu'il s'agit de méthodes basées sur des échantillons et que certaines d'entre elles utilisent la simulation de Monte-Carlo qui nécessite de nombreuses simulations pour couvrir l'ensemble de l'espace des paramètres et donner des résultats relativement précis. Cela devient un problème si le modèle utilisé nécessite une durée importante pour la simulation. En pratique, lors de la mise en œuvre d'une analyse de sensibilité globale, un compromis entre précision et coût de calcul est pris en considération dans la technique d'échantillonnage et dans le choix de la méthode appropriée.

L'analyse de sensibilité garantit que chaque paramètre sélectionné est identifiable compte tenu de la structure du modèle lorsqu'il est considéré seul ; cependant, elle n'indique pas si la combinaison des paramètres sélectionnés est également identifiable ou non. L'identifiabilité des paramètres est assurée si les paramètres du modèle peuvent être déduits de manière unique à partir des données. La structure du modèle ainsi que les données disponibles conditionnent l'identifiabilité des paramètres. L'interaction qui pourrait exister entre les paramètres les plus influents pourrait rendre cette combinaison non identifiable. Par conséquent, il est nécessaire de quantifier l’identifiabilité et d’adapter si besoin le jeu de paramètres avant de lancer le calibrage. La méthode de calibrage est expliquée plus en détail dans ce chapitre.

L'analyse bayésienne est une approche automatisée basée sur les probabilités qui permet d'améliorer la fiabilité d’un modèle en affinant une fonction de densité de probabilité (PDF) d'un paramètre d'entrée en fonction des données mesurées. Les paramètres estimés à partir d'une approche bayésienne prennent la forme d'une distribution de probabilité qui permet de calculer la confiance dans ces estimations et de mener une propagation de l'incertitude. Il existe de nombreuses approches de calibrage bayésien. Elles peuvent être classées en approches dépendantes de la vraisemblance et indépendantes de la vraisemblance, également appelées calcul bayésien approché (ABC).

Dans ce chapitre, une revue approfondie de la littérature sur les sujets énumérés ci-dessus est effectuée.

1.1 Introduction

This chapter aims at providing a background about the presence of uncertainties inherited within the building energy models and the importance of tackling it for better and more reliable predictions. A brief description of the most used dynamic building energy models (DBEM) is provided. In order to account for the uncertainties in a model, it is important to specify their sources and to quantify them in accordance with the calibration methods that will take advantage of this quantification to calibrate DBEM.

Different ways to quantify uncertainties in the DBEM parameters depending on the characteristics of these parameters can be found in the literature. A short review on different uncertainty quantification ideas adapted by the researchers in the field is presented.

To undergo calibration, it is important to select the most influential parameters and to estimate the degree of interaction between them. Different sensitivity analysis methods having different characteristics exist in literature. A brief description of different methods and their application in the field on DBEM is provided. Sensitivity-based identifiability analysis is also proposed to be used in the field of building energy efficiency.

Different calibration methods classified as manual and automated methods have been used in this field. An overview of their application and characteristics is provided especially the Bayesian calibration methods.

1.2 Building energy models

The use of BEM has been increasing in the last years to aid optimise and invest in the scenarios and designs that have the greatest effect on the buildings efficiency and occupants comfort. Those tools have evolved through three generations starting from "simplified methods" where many simplifying assumptions were considered and the results were very indicative. This was the first generation of BEM. The second generation arose when the dynamic behaviour of the buildings started to be slightly considered rather than just carrying out steady state calculations which do not reflect the real state of the buildings. The software tools that are currently widely spread and utilized are the third generation of the BEM where the dynamic behaviours became much easier to compute with the developments in the computational

technologies. Several simulation tools are used internationally as decision-making aid tools for the building designs and system implementations.

Pleiades is a dynamic building energy simulation tool that has been developed at Mines ParisTech for more than thirty years, originally for the bioclimatic design of buildings, and later to account for their environmental impacts. The software enables different types of calculations including regulatory verification, sizing of equipment, indoor air quality, and statistical analysis. Dynamic thermal and energy simulation are also performed via COMFIE calculation engine. COMFIE is based on a multizone finite volumes modelling with model reduction technique used to significantly reduce the simulation time while offering the same precise simulation results with the corresponding sensitivity to the design parameters. This is very essential especially when many simulation runs are required as it is the case when uncertainty analysis and calibration techniques are applied. It calculates the heating and cooling load, humidity and temperature in each zone of the building while accounting for heat transfer between zones. This model passed several validation tests (Peuportier 2005) on different case studies and proved to be precise whether compared to experimental measurements (Munaretto et al., 2017) or other international building energy models like EnergyPlus, TRNSYS, etc. (Brun et al., 2009).

Uncertainty analysis has been recorded extensively in literature in application on these two tools as they are internationally the most used ones (especially EnergyPlus). Brun et al. (2009) conducted a comparative study between the five most used BEM in France in the domain of buildings energy (Energy Plus, TRNSYS, Pleiades + COMFIE, CoDyBa, PHPP) and clear agreement in the results were shown between all the tools.

However, even though these programs can be sophisticated in estimating the behaviour of a building, they still suffer from discrepancy between the simulation results and the experimental measurements that can be revealed after carrying out validation tests. This is due to the uncertainties associated with the simulation which are classified as different sources of uncertainties in numerical models as listed earlier. One very important source is the parameters uncertainty that is associated with the parameters that describe the case study on which the simulation is carried out. In the building context, this source of uncertainty is a critical issue. Even if we are certain about a parameter, it is not guaranteed that it will remain constant over time as the building is subject to different climatic and other changes. As stated by Tian et al. (2018), different types of uncertain parameters can be classified in the building context: weather

data, building envelope, HVAC systems, and occupants behaviour. Those can be generally classified as dynamic inputs or static inputs. The dynamic inputs which are also called variable inputs are those that vary during the simulation with time such as the meteorological data: the outside temperature, the humidity, wind speed, etc., The static inputs which are also called the model parameters are those that have constant values throughout the simulation. They can be the parameters that describe the materials properties (assuming they do not vary), building geometry, and other specification of the energy systems inside the building.

The global purpose of this thesis is to enhance the accuracy in the BEM predictions. Calibration and uncertainty propagation analyses allow to estimate the model parameters and provide a degree of confidence in its predictions. We will focus solely on the COMFIE model as it has been extensively validated. Moreover, it has been used for different uncertainty analyses, optimisations, and calibration methods. The access to the code allows us to perform modifications based on our needs in the software simulation settings.

1.3 Parameters uncertainty

Our main focus in this thesis is to account for one source of numerical models uncertainties which is related to the indication of the parameters that are the basis to carry out the simulation. Normally these parameters have to describe precisely the case under study to be able to accurately predict the observed energy consumption or to forecast the future behaviour of a building that is being designed. However, in practice most of these parameters are not precisely known and cannot be assigned a specific value due to various sources of uncertainty. Different studies aimed at collecting as much experimental data as possible of the parameters usually used in building simulation. For example, Clarke and al. (1991) provided a comprehensive review of the available data worldwide regarding the thermo-physical properties of materials used in building simulation. Such studies are the basis on which parameter uncertainty quantification depends. Uncertainty quantification is an essential step in the uncertainty analysis methodology and the more realistic those quantifications and distributions are the more confident we can be about the simulation output uncertainty bounds.

Different studies focused on uncertainty quantifications of the building simulation parameters. In these studies, minimum and maximum bounds and PDFs (probability distribution functions) are fitted to the parameters based on empirical data from official databases and researches. Macdonald (2002) applied uncertainty analysis preceded by

uncertainty quantification of different types of uncertain building simulation parameters (thermo-physical properties, surface properties, internal gains, and infiltration rate). He based his work on different empirical data found in the literature and his quantifications were used in subsequent uncertainty analysis applications by many studies.

Sun (2014) also quantified different types of parameters using different databases (micro-climate parameters, materials properties, building envelope, and internal gains). Regarding the microclimate parameters, he used different models that estimate their values and quantified the uncertainties impacting the results of each model. He then used these quantifications to assign the uncertainty bounds to these models predictions. Lee et al. (2013) selected the seven most influential parameters indicated by a sensitivity analysis and fitted PDFs that describe their uncertainties. They based their work on different empirical data sets which describe the variability in the observations chosen from the literature for each parameter. They also used three goodness-of-fit tests to accept or reject the null hypothesis i.e. the distributions assigned to the parameters fit the observed data sufficiently with a 5 % significance level. The quantified parameters are the temperature setpoint, the chiller plant COP, internal gains, and outdoor temperatures. Normally, the degree of uncertainty of some parameters can be highly related to the case study in hand. If specific in-situ measurements are carried out to determine a parameter value (e.g. albedo on the site), the uncertainty weighting in this parameter could then be a function of how accurate these measurements are and this will be more realistic for the case study than relying on uncertainty bounds from the literature.

This section defines all the types of uncertain parameters in building energy simulation and provide a literature review on how each type is handled in uncertainty analysis methodology.

1.3.1 Building envelope

The building envelope combines the static parameters of the building including the thermo-physical properties of its materials and its surface properties such as the emissivity. Other uncertain parameters can be put under this group: the infiltration rate, thermal bridges, etc.

Even though, these are considered as static parameters having a constant values that does not change with time and are not stochastic in nature, they might end up having variability in

their values affected with time as they are subjected to different conditions through time such as the effect of the humidity on the thermo-physical properties of a material. The properties given by the manufacturer corresponding to the construction materials used in the building (e.g. thermo-physical properties) might be uncertain due to experimental measurements errors in the determination of these values. Moreover, the materials in use are not the same materials that are exposed to the experiments while evaluating their properties, and this adds another source of manufacturing uncertainty.

1.3.1.1 Infiltration rate

The air infiltration flowrate acts like a scenario that varies with time and is not considered as a static parameter as it is globally the case for all the parameters that are grouped under the "building envelope". That is due to the fact that air infiltration is function of the pressure difference between the interior and the exterior of the building that is in its turn dependent on the varying external weather conditions. However, in practice, the infiltration rate is generally taken as an average annual value based on the case study at hand calculated from correlations present in standards. One can also define a constant value for the building airtightness from which the infiltration rate can be estimated and thus, the airtightness can be calibrated. However, the infiltration estimation requires accurate knowledge of the wind pressure distribution over the building envelope and the complexity and uncertainty in determining this pressure distribution (wind pressure coefficient) lead to additional uncertainties in the infiltration rate evaluation.

Heo et al. (2012) collected infiltration rate measured data from 10 naturally ventilated office buildings recorded at a pressure of 50 Pa and compared the values with those recommended in the standards ATTMA and CIBSE. They found that the measured minimum and maximum infiltration rate data were higher than the ranges found in the standards and they quantified the minimum and maximum values as 0.10 and 1.25 h^{-1} (number of air changes per hour). For the calibration approach that they aimed to execute, they assigned a triangular distribution of this parameter as a prior with the quantified minimum and maximum values.

For the uncertainty analysis carried out by Munaretto (2014), the correlation of the standard EN 12831 was used:

$$Q_{infilt} = 2V_i n_{50} e_i \epsilon_i \quad (1.1)$$

where V_{i} is the volume of the heated space, n_{50} is the air change rate in h^{-1} of the building at pressure difference $50\,Pa$ between the building interior and exterior. e_i is the exposure coefficient, and ϵ_i is a correction factor that accounts for the increase in the wind speed with the height of the heated space from the ground level.

His case study was a house built within the INCAS platform of INES in the "Le Bourget-du-Lac". Following the standard, the exposure coefficient chosen e_i was 0.05 as the site of the case study is unsheltered, and the correction factor of the wind speed ϵ_i was taken 1 as recommended by the standard since the house height is less than 10 m. The air change rate n_{50} was taken from the in-situ measurements that was done by CETE of Lyon in February 2010 that yielded a value of $0.26\ h^{-1}$ with a pressure difference of 50 *Pa*. As a result, with a volume of heated space of $271\,m^3$, the infiltration rate used in the analysis was calculated to be $7.05\,m^3/hr$ ($0.03\ Vol/hr$). This value was considered for all the rooms of the house except for the attic and crawl space where he used values of 3 and 1 Vol/hr respectively (not from the correlation in the standard). As for the uncertainty bounds, a range of $\pm 10\,\%$ was added. However, regarding the attic and crawl space, higher degrees of uncertainty ($\pm 1 vol/hr$ and $\pm 0.5\ vol/hr$) were respectively added as they are highly uncertain. In the model, he chose not to include the infiltration rate when the mechanical ventilation was on. After all, the infiltration was not found to be an influential parameter after applying the sensitivity analysis and was chosen to be discarded and not used in the uncertainty analysis.

In the thesis of Robillart (2015), the same case study was used. The aim was to apply an inverse uncertainty analysis rather than an uncertainty propagation that was applied by Munaretto (2014). The same value $0.03\ Vol/hr$ for the infiltration rate was also considered. After applying the sensitivity analysis of Morris, it was found that the infiltration rate was not an influential parameter as the mechanical ventilation rate took over it and was calibrated instead.

Booth et al. (2012) applied blower door tests on four flats in their case study to measure the infiltration rate at 50 *Pa*. They compared their measurements with the CIBSE guidelines and found that in their case, the lower bound is much more certain than the upper bound as 75 % of their measurements data were closer to the lower bound recommended by the standard. Accordingly, in their Bayesian calibration framework, they assigned a prior Frechet distribution with a short-left tail where the values are more certain and a long right tail where the values are uncertain.

1.3.1.2 Albedo

The fraction of the incident radiation that is reflected from a surface (which is in this case the ground surrounding the building) is called the albedo. The measurement of the albedo is not an easy task and is subjected to some degrees of uncertainties. One of the reasons behind this is that the albedo-meter will receive different sources of diffuse radiation and not only the direct radiation incident from the sun.

Munaretto (2014) found that this is an influential parameter by the aid of the sensitivity analysis that he carried out, and it was included in the uncertainty analysis. An albedo-meter was placed on the soil surrounding the house and the measurements indicated a value of 0.35. They assumed it to remain constant the whole year despite the different conditions that may affect its value. In the uncertainty analysis, they have provided a uniform distribution for this parameter having 0.3 as a lower bound and 0.4 as an upper bound.

Thevenard and Haddad (2006) listed some of the albedo values in the absence of snow found in the literature and also generated models to estimate it in the presence of snow. From these values, Silva and Ghisi (2014) assigned a triangular distribution for the albedo parameter bounded by 0.13 & 0.26 in their uncertainty analysis that was applied to EnergyPlus model.

Sun et al. (2014) in their aim to quantify the uncertain microclimatic parameters of building energy simulation, have used the meso-scale model equation that calculates the ground reflectance based on the pervious and impervious road compositions (road solar reflectance and area fraction). The range of each parameter involved in the equation was collected from a global data set available in the literature and modeled the uncertainty of each with a uniform distribution. Then, following a Monte Carlo sampling approach, where particles were sampled from the uniform distributions, the meso-scale model equation was evaluated at each sample. This propagation concluded a distribution function that is centered on 0.25 and bounded between 0.05 and 0.4 for “terrain city”. These values do not work in the presence of snow. They based their uncertainty quantification for the albedo in the presence of snow on a literature review.

In his thesis, Robillart (2015) considered modelling the prior of the albedo parameter for calibration with a uniform distribution ranging from 0.28 to 0.42.

1.3.1.3 Thermo-physical properties

Not only the errors accompanied with the measurements impose uncertainties in the values estimated for the material thermo-physical properties, but also the moisture content, temperature variation, and ageing also change the base value of these properties and adds eventually a degree of uncertainty in using the base value Macdonald (2002).

Macdonald (2002) quantified the uncertainties of the conductivity, density, and specific heat of different materials used in the building construction based on the data available worldwide of the thermo-physical properties that are used in building simulation as collected in the report of Clarke et al. (1991). Actually, four classes of materials were held in the study (impermeable which has 0 % moisture content, non-hygroscopic which was assigned a 1 % moisture content, inorganic-porous which was assigned a 4 % moisture content, organic-hygroscopic which was assigned a 7 % moisture content). The material temperature variation was assumed to be 10 K in those uncertainty quantifications. He assigned a 5 % uncertainty on the measurement of the conductivity. The levels of moisture content affect the conductivity value and add uncertainty degrees of 5 %, 15 % and 25 % respectively for the aforementioned materials classes excluding the impermeable materials. The assumed temperature variation increases the uncertainty by 5 %. The materials density is only affected by the moisture content and not by the temperature change. The uncertainties caused by the moisture content are respectively 13 %, 4 %, and 11 %. An additional 1 % point of uncertainty is assigned to account for the measurement errors. The specific heat capacity was assigned the following respective uncertainty levels 4 %, 19 %, and 8 % caused by the assumed moisture content for these materials with additional 10 % point caused by the 10 *K* change in temperature, and a 12.25 % uncertainty level due to the measurements. Based on these considerations and quantifications, an average value and a standard deviation that describe the uncertainty impacting 36 classes of materials in the available data were assigned. These uncertainty bounds have been the basis of numerous uncertainty analysis conducted in building simulation.

Domínguez-Muñoz et al. (2010a) also quantified the uncertainties in the conductivity of insulation materials used in building simulation. They based their study on an extensive data set that includes products from different European manufacturers. Figure 1.1 adapted from Domínguez-Muñoz et al. (2010a) shows the minimal and maximal conductivities of materials and group of materials resulting from their quantification study.

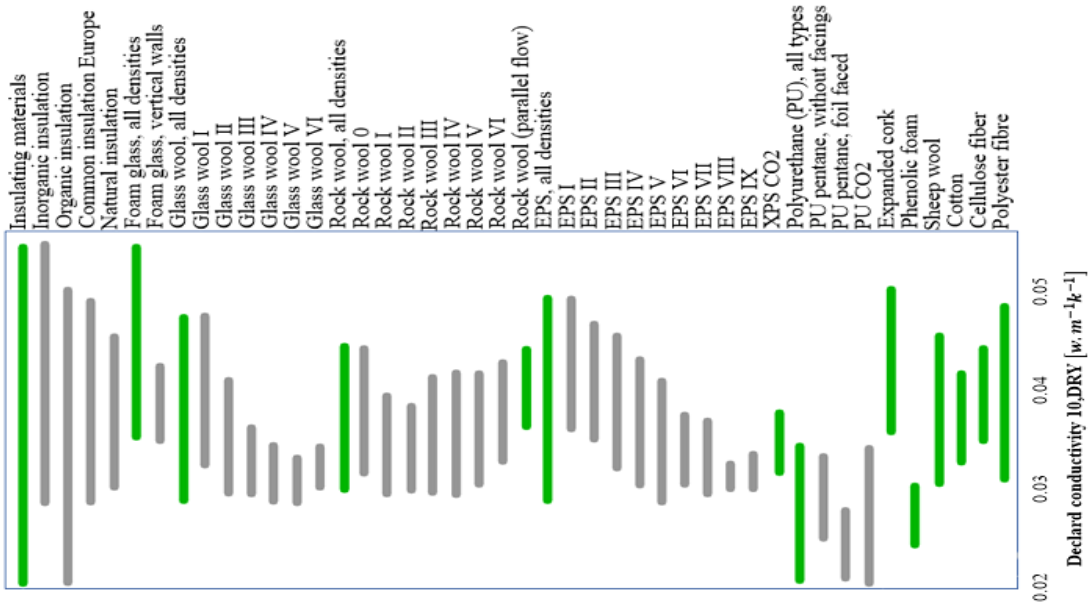

Figure 1.1: Max & min declared conductivities of materials and groups of materials (adapted from Domínguez-Muñoz et al. (2010a))

Munaretto (2014) assigned the thermo-physical properties of the materials in the case study of house IBB in the INCAS platform based on specific in-situ measurements taken by the CEA and listed in the (SIMINTHEC report). Some of the variations around these specified measurements for the uncertainty analysis were taken from the standard deviation pointed in Macdonald (2002). This was the methodology followed for both the conductivity and the volumetric heat capacity. When uncertainty bound was not found, ± 5 % was then considered. Among the influential parameters identified and propagated through uncertainty analysis there has been two thermo-physical properties: the conductivity of the wall polystyrene and the volumetric heat capacity of the concrete with means 0.03 W/mK and 2116 KJ/m^3K, and standard deviations of 0.003 W/mK and 210 KJ/m^3K respectively. The values are taken from the specific analysis of the study case from the (SIMINTHEC report), However, the uncertainty of the conductivity was taken from Macdonald (2002). Robillart (2015) modelled the prior distributions of these two parameters based on the same mean and standard deviation values.

1.3.1.4 Convective and radiative heat transfer coefficients

The surface heat transfer phenomenon is in itself complex and this explains why there exist many different correlations to estimate the heat transfer coefficients that may give considerably different results. The convective and radiative heat transfer coefficients and their correlations are subjected to uncertainties. They can be related to the uncertainties in the

measurements, intermediate calculation, and modeling assumptions used to derive those correlations (Driscoll and Landrum 2004).

Sun (2014) gathered a number of convective heat transfer coefficient correlations found in the literature and showed how the uncertainty in computing the coefficient can yield to differences by a factor of 10 between the minimal and maximal values if computed by different correlations. He then quantified this uncertainty by fitting a bivariate kernel density estimator to the coefficients of all of these correlations. The reason behind using a bivariate kernel is that the coefficients of the correlation were found to be correlated. He fitted such distribution for the external wall, internal wall, floor, and ceiling convective heat transfer coefficients by gathering the appropriate correlations corresponding to each type.

For the external heat transfer coefficients, the correlation is a function of the wind speed which is also another source of uncertainty in this correlation. As an extension to this work, Sun et al. (2014) used a standard ASHRAE model that downscales the measured wind speed at a meteorological station to predict the local wind speed at a coarser (local) resolution at the case study. They quantified the uncertainty of this model by analysing the differences between its results and the results of a high-fidelity Community Land Model (CLM). The models were run on a global data set comprised of different characteristics and parameters (e.g. tall building district, high, medium and low-density urban areas) to include all sources of uncertainty in modelling the differences between the results. This statistical model allows to obtain a distribution of differences between the two models for each height needed. As a result, quantifying the uncertainty of the local wind speed was done by adding a difference distribution (corresponding to the specified height) to the predicted local wind speed estimated by the standard ASHRAE model.

Wit (2001) considered several semi-empirical convective heat transfer coefficient correlations in his analysis and quantified the uncertainties weighting on this parameter to have a lower and upper bound based on the correlation results and on how widely spread they are. For example, for vertical walls, the internal convective heat transfer coefficient was chosen to be bounded by the correlation that yields the relative minimum value and the correlation that yields the relative maximum value among all the correlations used in the study. For the case of horizontal wall where the heat flow is upward, he used the same bound as those used for the horizontal heat flow cases but multiplied by 1.2 to account for the fact that the coefficients of

upward heat flow are commonly higher by 0-30 % than those of horizontal heat flow. Table 1.1 shows the correlations used as uncertainty bounds in this study.

Table 1.1: Uncertainty bounds of convective heat transfer coefficient assigned by Wit (2001)

	Convective heat transfer coefficient			
Bound	Internal (vertical wall)	Internal (upward heat flow)	Internal (downward heat flow)	External
Lower	Alamdari and Hammond 1983	(Alamdari and Hammond 1983) *1.2	ASHRAE 1997	Ito et al. 1972
Upper	Li et al. 1987	(Li et al. 1987) *1.2	Min et al. 1956	Sharples 1984

Munaretto (2014) also quantified the uncertainties impacting the convective and radiative heat transfer coefficients based on deterministic values found in the literature. He estimated the standard deviation for each coefficient based on the variability in the values assigned for them in the literature. All the convective coefficients were included in this quantification analysis (vertical wall, horizontal, internal, external). The convective heat transfer coefficient of vertical walls was found to be influential and was included in the uncertainty analysis.

1.3.1.5 Surface absorptivity and emissivity

Macdonald (2002) also quantified those parameters for different materials based on the data collected from Clarke et al. (1991) by assigning a base value and a standard deviation for each. Silva and Ghisi (2014) built on these quantifications and assigned a 0.02, and 0.04 standard deviation for the emissivity and solar absorptivity respectively. They have fitted normal distributions for these two parameters centered on a base value each of which correspond to the ceiling or the wall. Munaretto (2014) assigned the absorptivity and emissivity of the surfaces (external mortar, internal painting, internal side of the roof, floor tiling) by estimation, thus the uncertainties impacting these values can be high. They have assigned a ± 15 % degree of uncertainties on the absorptivity and ± 5 % on the emissivity. He has specifically assigned different bounds for the internal surface of the roof under the photovoltaic solar panels present in the house of his case study (± 50 % for the absorptivity and ± 20 % for the emissivity).

1.3.1.6 Thermal bridges

The thermal bridges effect on the building behaviour is dependent on the quality of the design and building construction. Low efficiency building can result if the designer's choice induce an interruption of insulation (e.g. internal facade insulation interrupted by a floor) or if it was constructed by a bad workmanship. Thermal bridge catalogues comprise coefficients based on fixed dimensions and materials. However, the configurations on which the coefficient values are listed may be different than those in the case study. This necessitates choosing a best fitting configuration in the catalogue which adds uncertainty. A flexible way is to calculate the thermal bridge coefficients for the particular case study configuration at hand using numerical tools. We call it flexible since the exact case study details are considered. As stated in the international standard ISO 14683, the numerical tools generally yield an accuracy of $\pm 5\%$ in the estimation of the lineic thermal brigde coefficient if the configuration is well defined. On the contrary, even if the loss through thermal bridges calculation is unbiased, poor construction can still add uncertainties to this parameter.

Moon (2005) quantified the severity of the thermal bridge effect using the temperature factor based on measurements and calculated results from the simulation tool KOBRA for three buildings. The temperature factor is a measure of how severe the thermal bridge is with 0 being the worst (internal surface temperature equals the external air temperature) case and 1 being the best case (internal surface temperature equals the internal air temperature). He focused on the thermal bridges at cavity wall corners. The idea of the analysis was to compare the measurements with the calculated results in order to quantify the uncertainty impacting the calculated values due to construction quality. No quantification about the thermal bridge coefficient was done in this study. Little studies have been carried out to quantify the uncertainties of this parameter. As stated by Wang et al. (2014), the thermal bridge effect quantification should be conducted on building specific bases.

Munaretto (2014) calculated the values of the thermal bridges using the software (TRISCO). They were calculated following two different conventions and used the differences in the results of each calculated coefficient to correspondingly set uncertainty bounds which at last were found to be varying between 10 and 30%. These variations were used in the sensitivity analysis but, in this careful design (external insulation avoiding interruption of insulation), were not found to be influential and were discarded in the uncertainty analysis.

1.3.2 HVAC systems

The HVAC systems performance also suffers from several sources of uncertainties. The uncertainties in the specifications related to the building design propagates to the HVAC system design. Ageing, maintenance and usual wear and tear also change the behaviour of these systems and consequently make their operation uncertain with time. The specifications provided by the manufacturer can also be misleading as normally the conditions on which the HVAC system will operate are different from those on which the specifications were identified.

Uncertainties are larger in the case of boilers, and even more for heat pumps, because of efficiency variations in terms of temperatures and heating loads. The same is true also for cooling systems. In the case study of Munaretto (2014), uncertainty intervals of $\pm 5\ \%$ were added to the measured nominal value of 1200 *W* heating power provided by an electrical resistance.

1.3.2.1 Internal gains

The uncertainty in identifying this parameter is related to the occupants' usage of appliances (domestic appliances, office appliances etc.) which is related to the occupants' behaviour. On the other hand, if the internal heat gain provided by the monitoring equipment is measured, therefore the uncertainty level is small. As stated by Parslo and Hejab (1992), this could be less than 1 %. Munaretto (2014) identified the values of the heat gain from the monitoring equipment in the house and considered $\pm 5\ \%$ uncertainty of the parameter in the uncertainty analysis. The total measured heat gain was 208 *W*. Robillart (2015) modelled a normal prior distribution for this parameter based on the same values. Normally, these internal heat gains can be very uncertain in the presence of occupants. Different studies have been carried out to account for this degree of uncertainty and to assign uncertainty intervals depending on the type of the building (whether it is e.g. an office or a residential building). In the study case of both Robillart and Munaretto, internal gain values were not uncertain as there were no occupants in the house during the experimental campaign.

Vorger et al. (2014) developed a statistical based occupancy model that simulates the presence/absence and activities of occupants during the day for residential and office buildings. Uncertainty propagation of the parameters of this model and the BEM allows for uncertainty analysis that considers the stochastic behaviour of the occupants. For calibration, this model is

not required since instead, real data concerning the occupancy behaviour sould be available and used. However, using the calibrated model to predict and analyse the uncertainty of the building performance in the future could require such stochastic occupancy model.

1.3.2.2 Temperature setpoint

Heo et al. (2012) used the Morris sensitivity analysis and deduced five influential parameters among which was the temperature setpoint. The temperature setpoint was 22 °C and a $\pm$ 2 $°C$ uncertainty was considered. They modeled this interval with a triangular distribution topping on 22°C and the resulted calibration posterior was found to be shifted a little towards a mean of 21 °C. The same interval was also adapted by McDonald and Strachan 2001 in their uncertainty analysis where the base value was 20 °C. Munaretto (2014) set a $\pm$ 0.5 $°C$ interval for this parameter but it was not included in his uncertainty analysis because the heating power was not controlled by a setpoint in the experimental protocol.

1.3.2.3 Mechanical ventilation

Munaretto (2014) used the measurements applied by CEA to estimate the mechanical ventilation flowrate in the building and found that the measured flowrate differs from that indicated by the manufacturer of the ventilator. Due to the different scenarios applied in the case study, two ventilation rates were considered depending on each scenario. The values measured by the CEA in their study indicated a nominal ventilation flow rate $Q_{v,nom1} = 110\,\mathrm{m}^3/\mathrm{hr}$ for scenarios 1,4,5 and $Q_{v,nom1} = 160\,\mathrm{m}^3/\mathrm{hr}$ for scenario 6. These values were assigned $\pm 15\,\%$ uncertainty.

1.4 Sensitivity analysis

Normally in the calibration process, after having collected and processed the data, a subset of the whole parameters set is chosen to be estimated from the measurements. This brings the need for sensitivity analysis (SA) which is often used as a basis to the uncertainty analysis. In the context of performing inverse uncertainty analysis as will be discussed in the following chapters, the computational demand becomes expensive if the number of parameters involved is too large. Thus, it is important to analyse the effect of all the model input parameters on the model outputs to perform the uncertainty analysis on those that are more influential and to

discard the parameters whose variations do not affect the model's output. In this context, discarding means to assign a constant value during the uncertainty analysis.

Sensitivity analysis is a statistical method that makes it possible to simplify or better understand numerical models by classifying and ranking uncertain factors according to their influence on the output uncertainty. Sensitivity indices are calculated for each factor in the model. The greater the value of the sensitivity index, the more influential the parameter. There are a large number of sensitivity analysis methods and the choice of one or the other must be made according to the objectives of the study. For a detailed description of various methods and their applications, the reader is referred to Pannier et al. (2018). In this section, a brief summary on popular methods is presented.

The general steps followed while performing the sensitivity analysis can be arranged as follows.

1. Define what are the model outputs that are of an interest.
2. Choose the parameters that need to be included in the analysis (this might be based on a previous experience and knowledge about some uninfluential parameters so as not to be included).
3. Define a distribution function for each input parameter with a known mean and variance (this could be done based on a previous experience).
4. Define a sampling technique to sample realisation from the parameters distributions.
5. Choose the most appropriate sensitivity analysis.
6. Carry out the selected sensitivity analysis on the sampled realisations from the input parameters distributions.
    a. Evaluate the model *N* times based on the samples taken from the distributions and on the simulation runs required for the selected SA.
    b. Analyse the model outputs and draw conclusions regarding the input parameters (the procedure of analysing the outputs is dependent on the method used).

These steps correspond to variance-based methods. They do not apply for other SA methods. For example, the local methods do not require an identification of the distributions for the input parameters as will be discussed in the following section.

There are different sampling techniques that could be adapted. The simplest one is the random Monte Carlo sampling where samples are taken randomly from the distribution. A more

effective sampling technique is the Latin hypercube sampling (LHS) which divides the distribution into intervals of equal probabilities and draws randomly one sample from each interval. This ensures that the samples generated cover all the distribution space which may not be accomplished in the random Monte Carlo technique. Quasi-random Monte Carlo is a sampling technique adapted by Sobol in his SA method called (Sobol indices). Different techniques that relate to a specific SA method should be adapted. For example, the FAST method follows a unique way of sampling that is sampling in the frequency domain and applying a Fourier transformation to the input parameters.

The sensitivity analysis methods can be classified into three different categories (Saltelli, 2008) though it is not the only way to consider classifying these methods.

1.4.1 Local sensitivity analysis

This method is a one-at-a-time method (OAT) since the approach is to change one factor at a time and undergo the simulation to detect the change in the simulation output caused purely by the perturbed factor. This is considered one of the simplest sensitivity methods and is used extensively in the literature (Spitz et al. 2012). It is categorised with the qualitative methods as it only gives a qualitative information about the importance of each parameter.

Spitz et al. (2012) applied the local sensitivity analysis in order to evaluate the uncertainty of building simulation applied to the INCAS experimental platform of INES. Sensitivity indices were calculated for 139 input parameters. They varied the parameters equally with $+1$ % from their nominal values and used the mean air temperature simulation output as the variable with which the indices are calculated. Since the output is a time series, a sensitivity index is calculated for each timestep allowing to monitor the variation of the parameters effect on the simulation with time. For each parameter, a mean sensitivity index $S_{i,m}$ and standard deviation $S_{i,std}$ were calculated, and instead of using the mean value to indicate the importance of the parameters, the distance $\sqrt{{S_{i,m}}^2 + {S_{i,std}}^2}$ was used. It was chosen because it detects the degree of variation which might be high for some parameters with low means indicating that they are influential in contrary to what the mean sensitivity index tells. Seventeen parameters grouped into ten different families comprising those parameters that have the same effect, were found to be the most influential. This grouping was possible through a correlation analysis applied on the parameters sensitivity mean and standard deviation indices. Ten parameters which mimic

the uncertainty caused by all the 17 parameters were chosen for the uncertainty analysis. The first four influential groups were the power of the electric heating, heat exchanger efficiency, conduction of insulation that includes its thickness and conductivity, and the internal gains.

1.4.2 Screening methods

A widely used screening method is Morris' method (Morris 1991). It is categorised as a one-at-a-time method where, for each model run, only one input parameter is changed. It is more accurate than local methods as it accounts for non-linearities and interactions between parameters where the sensitivity of a parameter may depend on the values of other parameters. For instance, in the case of constant temperature heating, the influence of thermal mass is negligible if solar or internal gains are small, but high in the case of a large area of well exposed glazing. Moreover, they are computationally less expensive than global methods. What makes it preferred by researchers in the context of buildings, is its ease of use in addition to the fact that it accounts for the interaction between the parameters.

Heo et al. 2012 used Simlab to carry out Morris' method to identify the four most important parameters; they were the intercept for windows opening (i.e. the constant in the logistic regression that computes the percentage of windows opeing based on the outdoor temperature), the indoor temperature, the infiltration rate, and the discharge coefficient.

Munaretto (2014) applied Morris' method to 153 parameters to deduce 11 most important parameters to propagate through the uncertainty analysis. He used 6 discretisation levels, with 30 repetitions requiring 4620 model runs. He classified the important parameters based on the distance d_j^*

$$d_j^* = \sqrt{\mu_j^{*2} + \sigma_i^2} \tag{1.2}$$

This method is widely used in the context of building simulation in the domain of uncertainty propagation and calibration where applying sensitivity analysis is an essential step.

1.4.3 Global sensitivity analysis

Contrary to the local methods, the global methods evaluate quantitatively the importance of each parameter over the whole input space taking into consideration the interaction between

parameters. Thus, they explore the entire parameter space. They are similar to Morris' method in terms that they are model independent approaches but they can be expensive computationally. The reason behind that is that they are sample based methods and some of them use Monte-Carlo simulation which requires many realisations to cover the entire parameter space and give relatively accurate results. This becomes a problem if the model at hand requires a significant amount of time for the simulation. In practice, when implementing global sensitivity analysis, a trade-off between accuracy and computational cost is taken into consideration in the sampling technique and in choosing the appropriate method.

1.4.3.1 Regression-based methods

Regression-based sensitivity analysis is one of the GSA methods. The main idea behind it is to fit a multivariate linear regression between the model output and the parameters by running a Monte-Carlo simulation. The methods vary depending on the complexity of the model. The SRC method requires a linear model with independent parameters to function properly, whereas partial correlation coefficient (PRC) method can handle correlated parameters. A widely used technique in regression methods is the forward stepwise technique (Tissot and Prieur 2012). Its idea is to add the parameters individually to the regression model starting with the most important parameter until no variable is significant anymore. The most important parameter is the one that provides the highest increases in the coeffient of determination R^2 or that provides the higest drop in the residual sum of squares. The advantage of regression methods is that they can be easily implemented and understood and they are computationally faster than the other global methods (Tian 2013). Different regression-based methods are reported for sensitivity analysis application in the building sector.

Domínguez-Muñoz et al. (2010b) used a regression-based method (SRC) to estimate the importance of different input parameters on the peak cooling load of an office in Malaga. The sensitivity analysis was applied using the uncertainty probability distributions on input parameters that are used for the uncertainty propagation according to the literature, theoretical considerations, and educated guesses as stated. The SRC associated with all the input parameters enabled ranking the parameters in decreasing order in terms of importance. Among the 20 uncertain parameters involved in the analysis, 8 were found to have an important effect on the peak load, two of which were related to the thermal inertia (internal thermal mass, and convective heat transfer coefficient between internal mass and the room air), and the rest were related to the internal gains, solar gains, and mechanical ventilation flowrate.

Yildiz et al. (2012) used the Standardised Rank Regression Coefficient (SRRC) sensitivity analysis to identify the parameters that greatly affect the annual cooling energy loads in low rise apartment buildings for Izmir (Turky) climate. The same analysis was done for the present climate and for climate scenarios expected in the 2020s, 2050s, 2080s. The same parameters were indicated to be the most important for the selected time periods with small differences in the order of importance of some parameters. 33 parameters which can be grouped into 10 different families were involved in the analysis and bounded according to commonly used data in the construction industry and regulations in Turkey. The most important families of parameters were estimated to be the natural ventilation, window area, and the solar factor of the glazing.

1.4.3.2 Variance-based methods

The idea behind the variance-based method is to decompose the model's output variance into a sum of variances that are caused by different input parameters. These variances are then transformed into indices to evaluate the parameters influence on the model results. For each parameter a first order index that explains its influence separately without accounting for the interaction with the other parameters, and a high order index that explains its interaction with other parameters can be calculated. A total sensitivity index that takes both into consideration can also be estimated for each parameter. The total sensitivity index is defined as the sum of all the sensitivity indices (separate and with interactions) involving the parameter in question. The parameters are introduced as probability distributions with known means and variances. These distributions can be taken from the literature or based on an expert's opinion.

Unlike Morris' method, the variance-based methods are quantitative methods that quantifies the amount by which a certain parameter is more influent than the other. In these methods, the model can be treated as a black box because it does not interfere with the method. However, the drawback of these methods is their computational cost because in order to explore the whole parameters space, they requires thousands of computations.

1.4.3.2.1 Sobol's method

Sobol's method implemented by Sobol and Shukman (1993) explains by how much each internal model parameter has contributed to the variance in the model output. It also includes the effect of interaction between the parameters and its contribution to the variance in the model output. Unlike the regression based methods, this method could be used no matter what and

how complex the model is. One main feature of Sobol's method is that it computes the first and the total sensitivity indices of each parameter, however, under a high computational cost required requiring $N(K+2)$ number of simulaitons where *N* is the size of samples and *K* is the number of parameters included.

Spitz et al. (2012) applied local sensitivity analysis on 139 uncertain parameters to deduce the most influential ones on the building energy performance. This was their first step to get useful information about what parameters can be fixed. Then, they used Sobol's method to quantify the influence of ten parameters selected from the local analysis. This required 6669 simulations. They used EnergyPlus as the modeling tool for their analysis. The analysis was applied to a single house test facility at INES in France (Le Bourget-du-Lac). They measured the local weather conditions on the site and made a weather file from these measurements.

1.4.3.2.2 Fourier amplitude sensitivity test (FAST)

The Fourier Amplitude Sensitivity Test (FAST) has been firstly proposed by Cukier et al. (1973) to study the uncertainties in the rate coefficients of chemical reactions. It is used to discretise the model output variance into partial variances corresponding to each parameter. It is based on transforming the *k*-dimensional parameter space into one dimensional space in the frequency domain using a transformation function.

One of its drawbacks is that it only evaluates the first order index of the parameters and it does not account for the quantification of the interactions between the parameters. If the indices of the parameters sum up to approximately 1, then this approach is sufficient as such summation indicates that the model has few interactions between its parameters. Otherwise, another approach has to be considered in order to quantify those interactions.The FAST method has been used in the building sector by several authors. Mechri et al. (2010) used the FAST method to study the effect of different parameters on the heating and cooling energy consumption in a building in Italy. They have included only six most important parameters in their perspective for the sensitivity analysis.

1.4.3.2.3 Extended FAST

Saltelli et al. (1999) extended the Fast method to be able to estimate the total index of each parameter (EFAST). The difference is that a high frequency is associated to the parameter under

investigation and low frequencies that have no limitations of interference are assigned to all other parameters. The idea is that the variance explained by all $x_{\sim i}$ (i.e. all parameters except x_i) parameters and their interactions are isolated in these low frequencies.

The benefit of this improvement compared to the classical FAST approach is that the interference is not a problem anymore and the difficulty in assigning the frequency set is avoided. However, it is less efficient computationally than the FAST approach because different sets of model evaluations are required to evaluate all the total effects leading to $N = k(2Mf + 1)$ where w_h is the highest frequency assigned.

Shen and Tzempelikos (2013) used the extended FAST (EFAST) method for their sensitivity analysis. They have used a thermal and lighting simulation model to estimate the daylighting and the energy performance of private office spaces situated in Philadelphia (USA). They included only seven parameters in the analysis which made the EFAST method a reasonable choice.

1.4.3.2.4 Random balance design (RBD-FAST)

Another method for computing the first order sensitivity indices was proposed by Tarantola et al. (2006). It is based on the random balance experimentations techniques elaborated by Satterthwaite (1959). The random balance design approach RBD avoids the difficulty of choosing appropriate frequencies as it is the case in the FAST method by assigning one single common frequency value to all the parameters. It is cheaper computationally than the FAST and EFAST methods and the number of realisations is independent from the number of parameters due to the common frequency used (Gatelli et al. 2009). Under one simulation done on all the sampled points, it is sufficient to calculate the first indices for all the parameters.

As discussed by Tarantola et al. (2006), the RBD method can yield first order indices estimates for the most important parameters at a higher accuracy compared to the Sobol method, with only 2000 simulations in the former against 10200 simulations in the latter. However, for the less important parameters, Sobol was found to provide better estimates, but this is useless in the context of influential parameters identification.

Goffart et al. (2017) used the RBD method to study the moisture effect on cooling energy demand and indoor air conditions for the climate of Singapore. The analysis was applied for 14 parameters that are related to the walls and building materials. They used a maximum

harmonic of M=10 which yielded a minimum number of simulations of 220. However, they chose 600 simulations to ensure full coverage of the input space. They had to study five different cases which then required a number of 3000 simulations.

1.4.3.2.5 Synthesized FAST

All of these introduced approaches require the input parameters to be independent, thus they cannot be accurately applied when the parameters are correlated. Xu and Gertner (2008) proposed a generalisation of the RBD method and added some modifications to account for the correlated parameters. Their method is called synthesised Fast. This method uses only one frequency assigned to all the parameters as in the RBD approach in order to avoid the interference and the aliasing effects. However, instead of randomly permuting the common variable s, they proposed to keep it periodic to generate the samples in the parameter space accordingly, and then to apply the permutation directly on the samples in the parameter space. This method as the previous ones has no restriction on the model used. However, it can also be applied with as many correlated/uncorrelated parameters as the user specifies.

Mara (2009) extended the RBD-FAST method to include the estimation of the total measures. They tested this algorithm using a mathematical function for which an analytical solution can be found for the sensitivity indices. The number of parameters was set to eight.. They tested the method with different sample sizes ranging from small samples of 128 to large samples of 2048 and compared the results with the analytical solution and with EFAST. Both methods were shown to yield approximately the same total indices accuracy at large sample sizes. At low sample sizes, the RBD-FAST was found to perform better in general and even accurately specifically in estimating the TS_i measures with low values.

1.4.3.2.6 Random balance design (RBD-Sobol)

RBD sampling technique can also be applied to the Sobol method in order to enhance its computational efficiency as stated by Mara and Joseph (2008). Goffart (2013), as a part of her thesis evaluating the uncertainties in the field of thermal and energy modeling of low consumption buildings, used the RBD Sobol method as a sensitivity analysis, applying the bootstrap resampling technique.

1.4.4 When to use what

Each model and project require a certain method of sensitivity analysis which fits better than the others. Each sensitivity analysis method has its own features and properties which might be suitable for a specific model but inappropriate to use in another. The choice of the method depends on several criteria (Saltelli, 2008):

- the correlations of the parameters;
- the model computational cost;
- the number of input parameters;
- the model characteristics (linear/non-linear) if a model depended method is used;
- the simplicity of the method;
- whether the method is model dependent or not;
- the need of qualitative or quantitative measures;
- whether including the interaction effects in the measures is needed or not.

Some methods share the same characteristics; thus, several methods may be suitable to the same problem. In the context of complex models which require a significant amount of time to run one simulation, the computational efficiency of the method is a primary concern.

One of the main drawbacks of the local sensitivity analysis is that the interactions between the parameters are not considered in the sensitivity measures. It also does not explore the entire space of the parameters, instead, it only explores its subspace around its base case. However, it is computationally more efficient than other more accurate methods. Thus, if the global sensitivity methods are prohibitive due to their computation time, the local method could be a good choice to provide some information about the model behaviour near the nominal values of its parameters, or in this case, a metamodel can be trained to replace the original expensive model. Another efficient use of the local methods is to apply them firstly on all the input parameters to detect those parameters that have negligible effect and then to apply a global method excluding those identified parameters. This enhances the global method computational efficiency as it is for some of them highly dependent on the number of parameters.

The regression global sensitivity analysis is model dependent. Some of these methods require the model to be linear and monotonic others require only monotonicity. If these conditions were met, it would be a good choice to use a regression-based method as they are

computationally less expensive than the variance-based methods. Otherwise, using these methods for a nonlinear non-monotone model enables the modeller to draw only limited conclusions from the sensitivity quantification.

Among the screening methods, Morris' sensitivity analysis is the most frequently used in the context of building energy models (Tian 2013). The main characteristic of this method and which makes it the preferred one for many studies on BEM calibration, is its computational cheapness compared to global methods, and that it estimates the interactions between the parameters. This method does not allow for the quantification of how much each parameter has contributed to the output variance. However, it can still rank the parameters with sufficient accuracy.

For more accurate quantification, the global methods are used and more specifically the variance based methods due to their relatively higher accuracy and to the ability some of this class methods to quantifying both the main and the interaction effects. Similar to Morris' method, they are model free approaches. However, they require a relatively large number of model computations.

Several variance-based methods are proposed in the literature among which the main difference regards the computational cost. Sobol and FAST methods are commonly used sensitivity methods (Tian 2013) due to their accuracy compared to other variance-based methods. Sobol, classical FAST, and the extended FAST computational cost is highly dependent upon the number of parameters with Sobol being the most accurate one. The classical FAST is much more efficient compared to the others but under the drawback of not evaluating the total effects. Thus, if the summation of the first indices estimated by the classical FAST is found to be considerably less than 1, the modeller should think of using another method that quantifies the interactions missed by the classical FAST. If the number of parameters is too large, and the model itself is expensive, Sobol and EFAST can be computationally burdensome inapplicable. In this case, the RBD-FAST could be the solution where the computational cost is equal to the sample size N (Tarantola et al., 2006). The main drawback though is that it only evaluates the first order indices. The improved version of the RBD-FAST estimates the total sensitivity indices and was proved to perform better than the EFAST with small sample sizes and as accurately with large sample sizes. However, it is computationally more expensive than the original RBD-FAST.

Those discussed variance-based methods do not handle correlated parameters. In the case of correlated parameters, the synthesised FAST, which builds on RBD (the original), could be suitable with the same number of simulations required as in RBD. Once again, if the sum of all the first order indices was close to 1, this method is then considered to be sufficient, otherwise, the modeller should think of using another method for more accurate measures.

1.5 Identifiability analysis

Sensitivity analysis ensures that each selected parameter is identifiable given the model structure when considered alone; however, it does not inform whether the combination of the selected parameters is also identifiable or not. Parameters identifiability is the concept of whether the model parameters can be uniquely inferred from the data. Model structure as well as the available data indicates whether parameters' identifiablity is attainable or not. Interaction that might exist between the most influential parameters could make this combination unidentifiable. Therefore, there is a need to quantify this and solve it before lanching calibration. In this section, based on literature, a brief description on different identifiability methods is provided.

Parameters identifiability is an indispensable model property for good calibration practice. In other words, if two model simulations $\eta(x, \theta_1)$, and $\eta(x, \theta_2)$ where x represents the uncalibrated parameters and θ is the parameter of the model η included for estimation, were found to be identical, then θ is considered identifiable only if the sets θ_1 and θ_2 are also identical.

$$\eta(x, \theta_1) = \eta(x, \theta_2) \quad \text{for} \quad \theta_1 \neq \theta_2 \quad \Longrightarrow \quad \text{unidentifiability} \tag{1.3}$$

Unidentifiabilities can be due to the model's structure itself no matter what the quantity or the quality of the available data is, and can be caused by the data insufficiency or the bias caused during the measurements campaign. The former is called structural unidentifiably and the latter is called practical unidentifiability after Raue et al. (2009).

1.5.1 Structural identifiability

As mentioned previously, structural identifiability is strictly related to the structure of the model and not to the system under investigation and its corresponding measurements. If a model

is structurally unidentifiable, then at least one of its parameter can vary without exposing any variation to the model output, or a set of parameters can change values correspondingly while maintaining an approximately fixed model output. If the correlation between the parameters is linear, then a fixed model output is attained with different combinations.

It is highly recommended to check for structural identifiability prior to performing parameters estimation. On the one hand, if it is overlooked, then there will be no way to identify the source of uncertainties in the parameters' estimates, whether it is due to the available data, the model structure, or both. This means that enhancing the data quality or quantity may or may not add any significant improvements to the parameter estimates. On the other hand, if structural identifiability is confirmed, then, it will be obvious that the uncertainty in the parameter estimates are mainly due to the quality of the data: its accuracy and the relevance of the experimental conditions to the parameters.

With a structurally identifiable model, assuming that the data is sufficient to ensure practical identifiability, the parameters can be determined from the data and unique solutions and probability distributions can be attained. However, if the model is structurally unidentifiable, then its parameters cannot be uniquely identified, meaning that numerous sets of parameter combinations will resemble the data which biases the estimation of the parameters and the final precision of the calibrated model. Having highlighted this, it should be emphasised that the purpose of calibration could be one of two objectives. One may want to estimate the real values of the model parameters; in that case, a thorough identifiability analysis needs to be executed. Alternatively, one may want to achieve good predictions with the model, without attaining a precise estimate of the parameters, however, the validity of the model outside the range of the training data may be limited. Thus, in both cases, structural identifiability analysis is indispensable to screen out the parameters that are highly unidentifiable to ensure an adequate performance of the inverse analysis.

Different structural identifiability approaches have been illustrated in the literature, some of which apply to linear models, and other apply to nonlinear models. Juricic (2020) used three approaches to asses the structural Identifiability of different RC models. She used Taylors series and Laplace expansion which only can be used with linear models and differential algebra which generalises to nonlinear models. It is out of the scope of this thesis to present those algorithms since the model used here is nonlinear and the differential algebra approach can become difficult to apply for complicated and large systems.

Another family of structural identifiability methods is called sensitivity based identifiability analysis and most often referred to as estimability analysis. Miao et al. (2011) classified it as a third class between the structural and practical identifiabilities. In fact, it is considered to be similar to structural identifiability. The concept is that after a sensitivity analysis is performed, a sensitivity-based identifiability method is conducted to examine the dependency of the sensitivity matrix columns and to re-rank the parameters taking into consideration not only the parameters' importance but also their interactions.

The orthogonalisation method was originally proposed by Yao et al. (2003). The idea is that it re-ranks the parameters while accounting for two aspects: the importance of the parameters, and the interaction between them. In this method, the columns of the temporal sensitivity matrix are projected on the column corresponding to the most influential parameter. The parameter associated with the column corresponding to the least magnitude after projection is considered the least interacting with the firstly selected parameter. This method scales well with increasing number of parameters except for the additional computational cost concerning the sensitivity analysis.

Another method introduced by Brun et al. (2001) is called the collinearity method. It is a combinatorial problem where subsets of the parameter set are chosen and assessed whether identifiable or not. The idea is that firstly, the most important parameters are selected and the rest are discarded from further analysis. A combinatorial analysis is then executed to the selected parameters and the largest size of parameters set needs to be identified by the user. For each parameter subset, a collinearity index is calculated from the corresponding sensitivity matrix.

Gábor et al. (2017) proposed to estimate the largest identifiable subset using combinatorial optimisation. They chose the variable neighbourhood search optimisation technique. This allows to estimate the maximum size of parameters that can be identifiable which depends on the minimum threshold. It is then possible to extract all identifiable and unidentifiable sets. It could be said that this work is an organisation of how to perform the collinearity method efficiently.

In summary, sensitivity-based and structural identifiability are very similar and the same information regarding the identifiability of the parameters can be extracted. The application of structural identifiability approaches to nonlinear model can become difficult with high

dimensional systems which is not the case in sensitivity based methods. In the field of buiding energy efficiency, structural identifiability methods to linear models have been applied, and up to our knowledge structural identifiability as well as sensitivity based methods were not applied to large non linear building energy models.

1.5.2 Practical identifiability

Unlike structural and sensitivity based identifiability, practical identifiability reveals if the parameters could be learnt from the available data even if they were confirmed to be structurally identifiable. Insufficient data may not supply enough information for some parameters and thus, affect their estimability rank. Beyond the quantity, the deteriorated quality of the data due to measurement errors and uncertainties which is commonly the case poses problems too. Contrary to structural identifiability, practical identifiability could only be checked after calibration.

One approach to estimate the practical identifiability is the profile likelihood. Lets define a parameter space consisting of two parameters $\{\theta_1, \theta_2\}$. Set θ_1 to a given value and then find the value of θ_2 that maximises the log-likelihood given that value of θ_1. Repeat this for different values of θ_1. This results in a function showing the maximum possible likelihood for each value of θ_1 and is called a profile likelihood. In this frame, practical non identifiability will show as a flatness in the likelihood.

This approach becomes burdensome with increasing numbers of parameters. In fact, it is one way of estimating the parameters of the model. It will then be inconvenient to run profile likelihood and then to start calibration because it will become computationally intensive. An alternative is to estimate the parameters following Bayesian approach and then apply identifiability analysis to the posteriors. That is, to check for the presence of flatness in the likelihoods of the estimated posteriors.

Yi et al. (2019) applied Bayesian inference and checked the identifiability of the estimated parameters by comparing the parameter range to its likelihood confidence interval. If the likelihood confidence interval (CI) of a parameter is close to its parameter range, the parameter would not be identifiable. They also used a biplot analysis to check for correlations between the parameters and they used principle component analysis (PCA) for dimensionality reduction. A biplot is used to visualise the parameters in two dimensions after reducing the dimension of the parameter space using PCA decomposition.

Juricic (2020) proporsed to use the KL-divergence metric to estimate the difference between the posteriors and the priors. This helps estimate how much the data is informative. If the posterior is identical to the prior, it means that the parameters could not be estimated from the data, either because the priors are so accurate that the data confirmed that the true distribution is actually the prior itself, or due to parameter un-identifiability. Thus comparing the posterior to the prior only gives information about how much was learnt from the data.

1.6 Calibration

Uncertainty analysis is applied to predictive processes in order to estimate how our knowledge and uncertainty about its driving factors would contribute to our confidence about its prediction and how uncertain it is. In our context, the predictive process is the building simulation model and the driving factors are the model parameters. Uncertainty analysis is a broad expression that comprises two different types of analysis: forward and inverse uncertainty quantification. The first type aims at propagating the uncertainty of the input parameters through the model to quantify their effect on the certainty of its outputs, that is estimating for example how the energy used or temperature profile in the building varies with the corresponding hundreds of parameter combinations. It is thus very important to accurately quantify the uncertainties of the parameters that are to be propagated.

The second type of uncertainty analysis is the inverse uncertainty quantification which, contrary to the previous type, aims at quantifying the parameters by indicating the values that yield simulation results consistent with in-situ case study experimental observations. This class is also called calibration. Models like COMFIE, EnergyPlus, etc. contain several hundred parameters each of which is associated with uncertainties. Eventually, to be able to rely on the results of these models and use them for renovation projects with a certain degree of confidence, the discrepancy between the observed data and the simulation results have to be minimised. This process of minimisation is called calibration. Calibration is a procedure to learn from data about the model structure. Given an appropriate model and sufficient data, the model parameters can be numerically inferred. Some calibration methods identify a unique solution, while others find a most probable set of solutions by sampling and resampling given the model at hand and the collected data.

Ahmad and Culp (2006) showed that uncalibrated simulations could have very high inaccuracy in predicting the energy use in a building, where the calculated total energy of the

building varied in the range of ±30 % compared to measured data. This inaccuracy significantly influences design choices especially in the case of green buildings that integrate passive heating and cooling. Thus, the main purpose of calibration is to improve the use of simulation in order to be able to predict the future real-world behaviour with higher accuracy.

Calibration methodologies used in the building sector can be broadly classified into manual and automated techniques (Coakley et al., 2014). The manual approaches do not include any kind of automated mathematical method to assist in the calibration process. On the contrary, automated approaches do not need any user intervention throughout the process, because they rely only on mathematical tools, although, they still require the user to tune their hyper-parameters prior to launching the process.

1.6.1 Manual calibration

The manual based calibration methods are the earliest techniques in the field. they are based on manually tuning and modifying the input parameters iteratively. This requires an expertise in the domain of building simulation so that the user will be capable of changing the parameters efficiently in a meaningful pattern.

Diamond et al. (1986) calibrated simulations using the DOE-2 program in the case of seven commercial building types under monthly and annual basis. They used the utility bills for an entire year and information about the buildings and their HVAC systems and operating schedules as observations to carry out the calibration.

O'Neill et al. (2011) manually tuned the input parameters to minimise the gap between measurements and simulated results after identifying the most influential parameters using a sensitivity analysis. They developed two models to be calibrated, one using EnergyPlus and the other using TRNSYS which enables to incorporate different global and local control sequences that could be hard to evaluate using EnergyPlus. They based their calibration on the electricity consumption. For the months where real weather data is available, the EnergyPlus calibrated model yielded results within ± 10 % from the recorded measured data whereas for the same months, the uncalibrated model results differed by 25 % to 40 %. The calibrated TRNSYS model was found to be less accurate due to the less accurate zoning as explained by the authors.

Royapoor and Roskilly (2015) manually calibrated EnergyPlus simulation using the ASHRAE guide 14-2002. This guide states that in order for the model to be considered

calibrated, the Mean Bias Error (MBE) values should be in the range of ± 10 %, and the Cumulative Variation of Root Mean Square Error (CVRMSE) should fall below 30 %. They applied their study on a 5-storey sandstone office and achieved the required calibration criteria. The calibration was based on the recorded electricity and gas consumption in 2012 and then used the actual measured zone temperatures to validate that the calibrated model can predict the zone temperatures with an accuracy of ± 1.5 °C for 99.5 % of the time and of ± 1 °C for 93.2 % of the time.

There are several manual methods that have been adapted in the past in the building sector. Reddy (2006) carried out a literature review on previous works done in this field using manual methods. Those methods may be based on some graphical advanced presentations to help in carrying out the discrepancy minimisation like the graphical statistical indices and the signature analysis. Since calibration required many data points for more accurate results (e.g. hourly instead of monthly data), it will become harder for the user of such a manual approach to identify the causes of difference among the input data, thus it becomes more complicated to specify the appropriate parameters that need to be tuned. Accordingly, these methods can be considered as the less practical ones in the field of BEM calibration (Coakley et al., 2014).

1.6.2 Non-Bayesian automated optimisation

The automated calibration techniques rely on mathematical and statistical approaches that tune the parameters automatically without the need of a user intervention. This makes the model calibration not only limited to experts in the field, but also users with little expertise can thus use those techniques. The automated methods can also be further classified into different approaches. In these methods, error metrics are also used to carry out the comparison. However, instead of manually tuning the parameters, optimisation methods based on numerical simulation are applied. These methods can then be used to perform calibration. For the optimisation, an objective function has to be defined. Usually in calibration application, the objective function is defined as a function of the difference between measured and simulated data which is evaluated using the error metric function defined.

Lavigne (2014) defined an objective function which is based on the monthly energy consumption and power demand. He worked with the DOE-2 software tool coupled with an optimisation algorithm (Maquardt–Levenberg nonlinear least squares method). He based his

calibration on the electrical energy consumption and power demand measurements and used two building study cases.

Yoon et al. (2011) used an optimisation-based calibration technique to calibrate the unknown input parameters involved in the estimation of the heat and mass transfer of a double-skin system. They used a lumped 1D model. They calibrated the most noticeable unknown parameters that cannot be estimated analytically such as the convective heat transfer coefficient, the air flow coefficient etc. They used a decoupled approach, where the airflow and the heat transfer models are interacting, each using the results of the other in the previous step. Each model had its own objective function, used to calibrate its own set of parameters.

Some of the optimisation methods are approaches that handle a range of probable values for each parameter and then these distributions propagate through the model using the Monte Carlo simulation. This ensures that the calibration results in a set of values as a solution instead of one single value Reddy (2006). However, the main objective of most of these methods is not to reduce the uncertainty: they are just meant to reduce the discrepancy between the simulation results and the observed data (Muehleisen and Bergerson 2016). In this context, the Bayesian based calibration is a better choice to both reduce the discrepancy and to quantify the uncertainties in the parameters and predictions. Different approaches can be followed in Bayesian calibration methodology and they are explained in the following section.

In section 1.6.3, a brief overview on the Bayesian algorithms corresponding to the two different families (likelihood-dependent/independent) is conducted. Note that there are many more approaches and improvements in the field that are not mentioned here. Only the methods that are well known in the field and are of an interest for the current thesis are presented.

1.6.3 Bayesian calibration

Bayesian analysis is an automated probabilistic based approach that allows enhancing our belief around a probability density function (PDF) of an input parameter given measured data. Bayesian approach maximises the likelihood that the model results are consistent with the measured data which is explained via probability distributions. This methodology is based on the Bayesian way of thinking. In Bayes' theorem, it is possible to update your belief about a certain parameter based on some conditional probability that relates this parameter to a certain event (measured data) using the following law:

$$p(\theta|Z) = \frac{p(\theta) \times p(Z|\theta)}{p(Z)} \tag{1.4}$$

The term $p(\theta)$ represents the so-called prior distribution of the parameter θ under investigation, which is based on a previous knowledge and expertise. The term $p(\theta|Z)$ is the conditional probability of the parameter theta given the data or certain events Z; it represents the posterior probability of the parameter θ which we aim at calibrating. The term $p(Z)$ is a normalising factor which is most of the time difficult to compute, thus, the relation might be reformulated as follows:

$$p(\theta|Z) \propto p(\theta) \times p(Z|\theta) \tag{1.5}$$

This makes the posterior distribution proportional to the multiplication of the likelihood function and the prior distribution. The link between the prior and the posterior is the likelihood function $p(Z|\theta)$, which is the conditional probability of the data observed given the parameter.

Kennedy and O'Hagan (2001) proposed a mathematical formulation that relates the measurement data with the model outputs as follows:

$$Z(x) = \eta(x, \theta) + \delta(x) + \varepsilon(x) \tag{1.6}$$

Before proceeding with the formulation, it should be distinguished between the model parameters and model inputs. The model inputs x such as the dry bulb temperature, and occupancy scenarios are not included in calibration. The model parameters θ that take fixed values during the model simulation such as the conductivities are considered for calibration.

This formulation states that no model is perfect, and even at the true value of the input parameter, the model will not exactly fit the measured data and there will stay some discrepancy that is referred to as model inadequacy (Kennedy and O'Hagan 2001). This is taken into account by the term $\delta(x)$. $\varepsilon(x)$ is the measurement errors occurred during collecting the measured data. $\eta(x, \theta)$ is the model output under the two specified input types and $Z(x)$ is the measured data under the conditions x. Accordingly, under known conditions x, this formulation accounts not only for the parameters uncertainties, but also for the other two types of errors that normally occur.

Different methodologies exist for Bayesian calibration, all of which follow the same concept as in equation (1.4). Globally, they are classified into two main groups: likelihood-independent and likelihood-dependent approaches. The main difference is that the former approximates the likelihood function with some metric and a defined threshold from which comes the name approximate Bayesian computation (ABC). ABC originated in the biological science department by Pritchard et al. (1999) and then has been improved and recently it is getting the attention of researchers in different fields. A likelihood-dependent approach was firstly introduced by Kennedy and O'Hagan (2002) and was firstly used in the building sector by Heo et al. (2012).

1.6.3.1 Likelihood-dependent approaches

If there exists an analytical form of the likelihood function, then it is possible to deduce the posterior distribution through Bayes' theorem: if the form of the posterior distribution is known, we can find the posterior probabilities of interest directly. However, if the posterior distribution form does not follow a convenient distribution, then it can be estimated by drawing samples from it (Turner and Van Zandt 2012). The difference between one Bayesian algorithm and the other is the sampling technique adopted.

One of the most used samplers in the domain is the Markov Chain Monte-Carlo (MCMC). At the beginning a sample (proposal) is drawn from the prior and simulated. Then another sample is drawn and is compared to the previous sample and is accepted or rejected with a certain probability (like the Metropolis-Hasting acceptance rejection rate). This is done iteratively until a sufficient number of samples that satisfy the posterior distribution are selected and they form an approximation of the posterior PDF.

MCMC quickly becomes inefficient with increasing number of parameters, and it highly depends on the proposal indicated by the user at the beginning (e.g. the value of the first sample generated) in which if they were inappropriately selected, biased samples could be generated. Another sampler which belongs to the same family is the Hamiltonian Monte Carlo (HMC) one. It is a gradient-based approach which selects new proposals by applying the Hamiltonian dynamics physics concept. It allows the sampler to select distant proposals which are still in the high probability region of the posterior and less correlated with the previous sample. The uncorrelated samples eventually allow the sampler to converge with smaller numbers of samples as compared to the previous approach. No-U-Turn sampler (NUTS) is an extension to

the HMC sampler. The difference is that it automatically selects the hyperparameters of the HMC algorithm which makes it possible to run HMC without user tuning.

Another family of samplers is the sequential Monte Carlo (SMC). Ching and Chen (2007) proposed the transitional MCMC (TMCMC) sampler which belongs to the SMC family. The general concept is that instead of directly sampling from the posterior, intermediate distributions could firstly be sampled before the posterior is reached hence the name sequential. At each step, the distribution to which the samples will belong will be closer to the posterior and further from the prior. At each iteration, the best samples of the previous one are resampled again and tuned to explore the parameter space instead of drawing new parameter values once again from the prior.

Minson et al. 2013 proposed in his algorithm CATMIP to run multiple Markov jumps for each chain at each iteration in which the number of jumps is automatically determined to stay within an acceptable acceptance rate instead of only doing one jump as in TMCMC. This modification makes the sampler less prone to be trapped in local minimums.

Another variant to the SMC sampler was followed in Adams et al. 2020. The main difference is related to the selection of the number of MCMC jumps identified at each iteration. The application of SMC samplers in the building sector is very rare. Most researchers tend to apply MCMC instead.

Heo et al. (2012) introduced the application of Bayesian calibration to the building sector where they used it to calibrate normative energy models applied on a UK campus building based on the measured gas energy consumption. They started with quantifying the uncertainty of the input parameters, and then they applied a sensitivity analysis based on Morris' method to identify the most influential parameters, then they used this approach to propagate the chosen parameters uncertainties and to refine their prior PDF into an appropriate posterior PDF. They used MCMC for this purpose.

Kim and Park (2016) applied two calibration techniques: a deterministic (optimisation) and a stochastic (BC Bayesian calibration). They used EnergyPlus as the simulation tool and applied the analysis on a 5-storey office building located in south Korea. Morris' screening method was applied in this study to point out the most influential parameters. They aimed also at investigating how calibration methods depend on the model quality and measurement errors.

Kristensen et al. (2017) used the same framework as Kennedy and O'Hagan (2001) to calibrate a simple ISO13790 calculation tool using the seven most influential parameters estimated from the Sobol sensitivity analysis method applied on 32 parameters. They used a monthly calibration resolution, and based their analysis on the annual energy use.

Lim and Zhai (2017) carried out a comparative study between different types of metamodels used in the Bayesian calibration framework applied to the building energy model EnergyPlus. Five different metamodels were tested (multiple linear regression model, neural network, support vector machine, multivariate adaptive regression splines, and Gaussian process emulators). They used the monthly average energy use intensity of electricity and gas estimated using the software based on true predefined values of the parameters that are to be calibrated so that they can validate the calibration results. They showed that all the metamodels used were able to yield posterior PDFs that predict the true values, the Gaussian process emulator being the most accurate metamodel but with the highest computational time.

Sokol et al. (2017) applied this framework on Urban Building Energy Model (UBEM). They split their data into a training data of 399 homes to predict the distributions of six uncertain parameters, and a test data of 2263 homes to validate those posterior PDFs. The approach is applied once using the measured monthly electricity and gas consumption and once using the annual data. They compared the models results when using the values of the posterior distributions with the deterministic values and showed that they better fit the observed data. They also clarified the importance of using appropriate time steps (hourly instead of monthy) in model simulations compared to aggregated data as the posteriors based on the monthly observations yielded higher accuracy.

As stated by Heo et al. (2015), using a Gaussian process meta model limits the method to aggregated energy data (i.e. monthly data). This is due to the resulting large sample size used to train the gaussian process model (which has a computational cost of $O(N^3)$). In order to make the Bayesian framework computationally applicable in the case of large datasets, Chong et al. (2017) proposed two modifications. Firstly, they proposed to reduce the sample size of the data by generating a subset that is representative of the original set by randomly sampling from it. They used a sample quality metric to measure how similar the subset is compared to the original set. The second modification is the use of NUTS-MCMC that converges faster in high dimensional problems. They applied their method on two case studies, TRNSYS model of water-cooled-chiller for a mixed use located in Singapore, and EnergyPlus model of the cooling

system of a 10-storey building in USA. They based their calibration on hourly measured data of the cooling energy consumption. They needed to calibrate 2 parameters in the TRNSYS model as they used Type 666 (at each time step 4 variable inputs and 2 parameters: COP of the chiller, and the chiller rated capacity are needed). They calibrated five parameters in the study using the EnergyPlus model.

As revealed by the results of Sokol et al. (2017), the use of smaller timesteps in the simulation such as hourly steps contributes to a more precise calibration. Thus, aggregating the data into yearly or monthly for the ease of calibration adds another source of bias to the inaccuracies already present in the approach. Recent studies focused on using hourly data (Chong and Menberg 2018; Kristensen et al. 2017; Menberg et al. 2017). In this thesis hourly data is used.

1.6.3.2 Approximate Bayesian computation (ABC)

In complex models, the likelihood functions might be intractable due to their complexity. Approximate Bayesian computation algorithms are appropriate in this case because they do not require such likelihood functions. It is based on approximating the likelihood function with a metric that computes the discrepancy between the observed measured data from experimental setups and the model outputs, and then deciding what parameter value to keep and what to discard until representative distributions of these parameters, that approximate the true posterior PDF, is obtained. This avoids the necessity of the likelihood function knowledge.

Generally, in complex systems, the data are highly dimensional (e.g. a temperature time series can consist of hundreds of data); this reduces the probability of generating model output that closely matches the observed data. This leads to a lower acceptance rate of input parameters and thus a computation inefficiency. A common solution to this problem is to summarise the measured data (Z) into a lower dimensional summary statistic $S(Z)$ which captures all the relevant information in (Z) and use it instead of the original data. Accordingly, the same summary statistic should be applied to the simulated outputs of the model (Y) in order to obtain $S(Y)$ and to perform the comparison.

This increases the efficiency of the computation as a low dimensional data is now considered, however it may introduce a bias in estimating the parameter values. This depends on how sufficient the summary statistic is with respect to the parameter (θ). If all the

information of (Z) about (θ) are captured by the summary statistics $S(Z)$, the estimation will be exactly the same as if all the data was introduced in the computation. But this is theoretical because in fact, selecting small number of sufficient summary statistics has been one of the major problems within ABC applications in different fields: only informative but insufficient statistics are often used in ABC applications (Harrison and Baker 2020). However, this is not very critical with BEM applications in the case of time series data, where the RMSE could be a sufficient summary statistic.

1.6.3.2.1 Rejection algorithm

The first ABC algorithm called rejection algorithm was proposed by Pritchard et al. (1999). They summarised their data set into three statistics $S(Z)$. There were four parameters (θ) to be calibrated. The rejection algorithm draws large number of samples from the priors and compares the corresponding model output with the measurements and accepts or rejects the draws based on a predefined tolerance value δ. If the difference between the model's output and the observed data is lower than this tolerance value, then the considered draw is accepted:

$$\begin{cases} S(Z) - S(Y) \leq \delta & \text{sample accepted} \\ S(Z) - S(Y) > \delta & \text{sample rejected} \end{cases} \quad (1.7)$$

The samples that respect the criteria are accepted under equal probabilities without taking into account the value of the difference itself whether it is very small or close to the tolerance. The parameter that yielded a smaller difference is more likely to be within the posterior than the one that yielded a higher difference even if both were below the indicated tolerance. Beaumont et al. (2002) refined the model proposed by Pritchard et al. (1999) by weighting the candidate θ_i according to the value of $|S(Z) - S(Y_i)|$ instead of accepting it with probability 1. They have also applied a local linear regression on the posterior to adjust the accepted parameters in order to weaken the effect of the difference between the observed data and the simulation output.

This regression adjustment of the posterior distribution after the termination of the ABC algorithm is called post-processing. Different ABC-post-processing techniques and approaches have been proposed by different authors. Some of these approaches are regression-based methodologies and some follow a different concept.

1.6.3.2.2 Hierarchical ABC

Hickerson et al. (2006) used the ABC framework of Beaumont et al. (2002) and developed a hierarchical model which is then called HABC. The idea behind this modification is that the prior distributions are function of some parameters (hyper-parameters) that are also inferred in the ABC framework by assigning prior distributions for them (hyper-priors). This allows to group the responses of the inferred parameters as a function of the inferred hyper-parameters. The probability of certain responses of the values of the calibrated parameters can be function of the variation of the values of the hyper-parameters. The hyper-parameter values are the highest level of the hierarchy on which the rest of the parameters are conditioned. This hierarchical modification is not only applied to rejection algorithm, however, it is also implemented to the updated versions of ABC as will be shown in the following sections.

1.6.3.2.3 ABC-Markov chain Monte Carlo (ABC-MCMC)

The rejection ABC is still computationally expensive since all the samples are drawn from the prior distribution, and if the posterior is significantly different from the prior, the great majority of drawn particles values will not yield acceptable simulation outputs, which contributes to a high rejection rate and low computation efficiency. Marjoram et al. (2003) proposed to use the Markov Chain approach described earlier in ABC. In this case, the Metropolis acceptance rejection probability is only applied when the discrepancy between the model output and the observed data is under the tolerance identified.

As explained by Sisson et al. (2007), the ABC-MCMC approach may get stuck in the regions where the probability of acceptance is very low which causes a high computation inefficiency especially if the proposal selected is bad. Besides the computational cost perspective, the MCMC algorithm usually suffers from poor mixing problems. That is to say that the algorithm may not explore the whole parameter space and sample from all the posterior regions. To account for this problem, Bortot et al. (2007) proposed to treat the tolerance as an additional parameter with a distribution so that at each step, along with sampling a particle from the distributions of the parameter space θ, a new value of the tolerance δ is also sampled from its distribution. This will allow to occasionally sample a high value of δ allowing to explore better the parameter space if the chain is stuck in a local minimum and to get out of areas of low probabilities. In this algorithm, a tight tolerance distribution can be used to keep the value

of δ closer to zero so that the number of particles accepted on a high tolerance is as low as possible.

1.6.3.2.4 ABC-sequential Monte Carlo (ABC-SMC)

Sisson et al. (2007) proposed the use of the SMC or "particle filter" concept in the approximate Bayesian computation context to solve the problem of convergence associated with the previous algorithms. The idea is similar to the SMC applied in the likelihood-dependent approaches. The criterion that defines the sequence of ditributions from the prior to the posterior is a sequence of decreasing tolerances. Instead of defining one tolerance value, a sequence is defined. At each iteration, the samples respecting the corresponding tolerance are kept. This enhances the computational efficiency of the approach compared to the previous ones. Sisson et al. (2007) proposed this method in combination with ABC under the name ABC-PRC (partial rejection control).

1.6.3.2.5 ABC-population Monte Carlo (ABC-PMC)

Beaumont et al. (2009) showed that the weighting ratio used to weight the particles in Sisson's approach was biased, and he proposed a different weighting scheme of the following form. The approach of Beaumont et al. (2009) named ABC-PMC (population Monte Carlo) is the same as the PRC approach but with corrected weighting ratios that avoid the bias presented in the PRC method.

The approach of Beaumont et al. (2009) combines the benefits of the basic rejection and MCMC algorithms. As in the MCMC, the parameter values are drawn from a distribution closer to the posterior instead of being drawn from the prior, and as in the rejection algorithm, this approach has no risk to be stuck in a region of low probability. The problem in this approach is in choosing the sequence of tolerance values through the iterations $(\delta_1, \delta_2 \ldots \delta_t)$, and in deciding when to stop the iterations. The sequence of decreasing tolerances can influence the accuracy of the obtained results and the computational efficiency. Moreover, if the tolerance value at the last iteration (δ_t) is too large, the final posterior will perform badly in estimating the true posterior, and on the other hand, the posterior resulted under using a very small tolerance value on the last iteration could have been approximately reached with a higher tolerance (less model runs).

1.6.3.2.6 Adaptive-PMC (APMC)

Lenormand et al. (2013) proposed a solution to this problem by determining the tolerance of each iteration based on the samples of the previous iteration. The tolerance (δ_t) is determined as the first quantile of the distances found in the previous iteration. They based their work on the PMC-ABC algorithm where they modified it with this criterion and different weighting function and they called their algorithm adaptive population Monte Carlo ABC (APMC).

The difference is that in the regular PMC-ABC, each particle in the previous iteration is perturbed until it is moved to a value that yields a lower discrepancy than a predefined tolerance at the current iteration and a new set of samples is formed that only contains the new particles. However, in this algorithm, each particle is perturbed only once and then weighted to be subjected to filtering after the indication of the appropriate tolerance.

Drovandi and Pettitt (2011) firstly suggested the idea of automatically computing the tolerance at each iteration based on an α-quantile. They used an MCMC kernel to avoid the bias present in Sisson et al. (2007) with a lower computational cost of $O(N)$. However, this kernel inherits the problem of particles duplication, that is the presence of the same sample multiple times in the sample set. APMC avoids this problem.

1.6.3.2.7 ABC in different fields

ABC methods have been extensively used in the field of biology and genetics. The first approximate Bayesian model was implemented by Pritchard et al. (1999) in the field of molecular biology as stated earlier. Then, Beaumont et al. (2002) extended the approach as explained in the same field where they based their tests on the models and data set analysed by Pritchard et al. (1999). Hickerson et al. (2006) used the framework of Beaumont et al. (2002) in the field of biology to test the simultaneous divergence between taxon pairs. Johnston et al. (2014) used Marjoram et al. (2003) framework of ABC (ABC-MCMC) to estimate the cell diffusivity and the cell proliferation to enhance the drugs design for the treatment of chronic wounds and different applications. Fan and Kubatko (2011) estimated the species tree topology and branch lengths using the ABC algorithm. They applied the same framework of Pritchard et al. (1999). However, they did not use a tolerance to reject or accept any of the sampled priors, instead, they simulated samples from the priors *N* times and recorded the distances and then retained the ones with the relatively smallest distances. It has also been used in different fields

like in image analysis, hydrological models, ecology etc. A summary of the fields in which ABC methods were implemented can be found in Sisson et al. (2018).

In the context of building energy models calibration, the likelihood-free inference (ABC framework) was firstly implemented by Robillart (2015). The aim of his thesis was to develop real time control strategies for electric load shifting in energy efficient buildings, so the simulation tool used was calibrated using the ABC-PMC approach. The DBEM was (COMFIE) which is the most widely used in France in the field of dynamic building energy simulation. The case study was a house of the INCAS platform of INES (Institute National de l'Énergie Solaire) located in Le Bourget-du-Lac. The experimental protocol included six different scenarios. For each scenario, several operating parameters were modified such as the heating set point, shutter state, mechanical ventilation air flow rate etc. Four of these scenarios were used in the sensitivity analysis and calibration, and two of them were used for validation. He evaluated the influence of 102 uncertain parameters using the Morris sensitivity analysis and chose the six most influential parameters related to HVAC system, heat gains, and materials properties. The prior distributions fit to these parameters were either normal or uniform. The difference between the simulated results and the measured data during calibration was calculated using the mean square distance:

$$\rho\left(T_i^{sim}, T^{mes}\right) = RMSE_i = \sqrt{\left(\sum_{k=1}^{N} \left(T_i^{sim}(k\Delta t) - T^{mes}(k\Delta t)\right)^2 / N\right)} \qquad (1.8)$$

where Δt is the simulation time step, and $T_i^{sim}(k\Delta t)$ and $T^{mes}(k\Delta t)$ are the simulated and measured temperatures at instant $k\Delta t$. He based the calibration on the temperature measurements rather than on the energy consumption data. The algorithm proposed by Beaumont et al. (2009) ABC-PMC was used as the calibration method in this thesis with tolerance varying from 20°C to 0.4°C at the last iteration. In each iteration, the number of particles that needs to be accepted under the specified tolerance was chosen to be 300. The model used in his thesis is a monozone model and not multizone which necessitated averaging the measured data over all the zones based on the areas corresponding to each zone to be able to apply the comparison with the model output which affects the accuracy of the measurements. The experimental campaign lasted for four months from January to the end of April and six different scenarios were applied to the house with each lasting for a certain period of time. The data collected for four of these scenarios was used for the sensitivity analysis and the calibration

methodology to identify the posterior distributions. The other two scenarios were used for validation. The comparison of the validation data (experimental scenarios 4 and 6 in his thesis) when simulated by the model before and after calibration showed some improvement. He explained the small decrease in RMSE in several ways. Firstly, the model considered in this study is a single-zone model whose precision is lower than that of a multizone model. Secondly, in order to compare the simulated temperatures with the measured temperatures, the latter were aggregated in proportion to the interior surfaces of the rooms in the building, also leading to a decrease in accuracy.

Zhu et al. (2020) applied the ABC approach based on Pritchard et al. 1999 framework with different posterior post-processing techniques to reduce the effect of high discrepancy threshold normally used in this algorithm. They applied and compared the performance of three different techniques: no post processing (Pritchard et al., 1999), ridge regression (Beaumont et al. 2002), neural network (Blum et al. 2010). They also used machine-learning methods to replace the detailed energy model to generate 100,000 samples from the posterior. The applied method as stated previously suffers from computational burden and robustness issues since all the samples are drawn from the prior distribution, and if the posterior is significantly different from the prior, the great majority of drawn samples will not yield acceptable simulation outputs, which contributes to a high rejection rate and low computation efficiency. ABC-PMC applied by Robillart (2015) overcomes this problem by sampling from a kernel closer to the posterior rather than sampling from the prior.

1.7 Conclusion

The use of BEM has been increasing in the past decades. Such simulation tools cannot be reliable if they are not validated with observed data. Generally, those models are built on simplifications and assumptions which could make the predictions inconsistent with the real behaviour even if the study case is assumed to be well specified. In practice, the case study cannot be defined in those models exactly as it is in reality due to the uncertainties in identifying the values of its parameters. Those uncertainties propagate through the models impacting their outputs and yielding inaccurate predictions.

Uncertainty propagation has been extensively applied in building simulation where it enables the user to quantify how much uncertain the simulation results are. Inversely, the calibration methodology aims at identifying the BEM parameters values using in-situ data by

tuning parameters and running the model. This tuning procedure can be done in different ways depending on the calibration method and its settings.

Different calibration methodologies in building simulation are found in the literature and are briefly reviewed in this chapter. Among these methods, we settled our choice on the Bayesian calibration technique since it accounts for uncertainties impacting the parameters and since the calibrated results are in the form of probability distributions functions for these parameters. This is advantageous against other calibration techniques that generate a unique value for each parameter. The application of this method requires fitting distributions (priors) to the input parameters in question and then they are updated through the algorithm to generate posterior distributions that are more likely to yield model predictions which are more consistent with observations.

Their dependence on the fitted priors necessitates to accurately identify those distributions. The uncertainty bounds impacting each parameter should be carefully evaluated based on empirical data and different standards found in literature. Some of the articles that focused their work on setting up these bounds for different parameters of building simulation are presented in this chapter.

Another essential step that also precedes calibration is the so called sensitivity analysis. The aim is to reduce as much as possible the number of considered parameters in the calibration process due to the computational burden. Moreover, including all the parameters caused what is called over parametrisation. This means that there will exist huge interactions allowing for different combinations that fit the measurements correctly but with no physical interpretation to the parameters estimates which might in turn affect the model accuracy. A brief overview on different sensitivity analysis methods available and used in building simulation is also provided in this chapter.

Even the most important parameters might have high interactions which makes it difficult to calibrate them at once. To this end, the identifiability analysis is an essential technique to quantify this issue. It enables to select the parameters by accounting to their importance and interactions at the same time. A brief description of this concept is conducted in this chapter.

In the following chapters, the described methods and concepts will be implemented in the calibration framework with real and synthetic data in an objective to study and enhance their application to building energy models.

Chapter 2

Sensitivity methods assessment

Since calibration is performed in a subset of the parameters, it is essential to rank the parameters correctly in terms of the most to the least important one. The most accurate sensitivity method is computationally intensive; therefore, it is not preferred in practice. This chapter focuses on different sensitivity methods available. A comparative analysis is conducted between the retained methods to assess their precision, robustness and computational efficiency.

Résumé du chapitre

Pour faciliter le calibrage, il est important de sélectionner les paramètres les plus influents de manière à réduire le coût de calcul. Une analyse de sensibilité classe les paramètres du plus influent au moins influent. Il existe différentes analyses de sensibilité qui diffèrent par leur efficacité et leur précision. Dans ce chapitre, Morris et RBD-FAST sont retenues de la littérature. Morris est une méthode de sensibilité très connue qui est largement utilisée dans le domaine des modèles énergétiques des bâtiments en raison de son efficacité de calcul et de sa précision suffisante. RBD-FAST est une méthode de sensibilité efficace sur le plan informatique qui a été récemment appliquée aux modèles énergétiques des bâtiments. Une comparaison détaillée est effectuée entre ces deux méthodes en termes de précision, de robustesse et d'efficacité de calcul. La méthode Sobol est considérée comme la méthode de référence dans ce chapitre.

Il est important de mentionner que la grandeur d'intérêt, qui est dans notre cas le profil de température, est découpée en pas de temps de 24 heures. C'est-à-dire que pour chaque période de 24 heures, l'erreur quadratique moyenne (RMSE) entre les mesures et les simulations est évaluée. Les indices de sensibilité sont alors calculés pour chaque paramètre à chaque pas de temps. Ceci génère une matrice d'indices de sensibilité ; chaque colonne correspond à un vecteur de sensibilité d'un paramètre distinct. Pour classer les paramètres, la norme euclidienne de chaque vecteur est calculée et les paramètres sont classés en fonction de ces valeurs.

Les critères sélectionnés pour la comparaison sont la précision des méthodes dans le classement précis des paramètres par rapport au classement par la méthode Sobol, leur robustesse dans le classement de tous les paramètres, ceux responsables de 95 % de la variance totale et ceux responsables de 90 % de la variance totale. De plus, leur efficacité de calcul est également prise en compte en évaluant leurs performances avec un nombre croissant d'évaluations de modèles. Les indicateurs utilisés sont principalement, le coefficient de corrélation de Pearson et le coefficient tau de Kendall. Les méthodes sont appliquées sur une étude de cas : une maison individuelle, dont le modèle comporte 113 paramètres.

La méthode de Morris montre une très bonne performance même avec un petit nombre de répétitions donnant des rangs approximativement similaires à la méthode de Sobol même sur les paramètres relativement peu influents. Elle est très efficace pour regrouper les paramètres les plus influents même si, à certains moments, elle ne classe pas exactement tous les paramètres

influents identifiés par la méthode de Sobol. Une certaine variation dans les rangs de certains paramètres influents est observée mais elle est négligeable. Cependant, le classement entre les troisième et quatrième paramètres est peu reproductible, ce qui n'est pas le cas avec RBD-FAST.

RBD-FAST montre une bonne performance dans le classement des paramètres responsables de 90 % de la variance totale estimée par la méthode de Sobol, en particulier pour les trois premiers paramètres. Ces paramètres sont systématiquement regroupés avec un nombre relativement faible d'évaluations de modèles par rapport à la méthode de Sobol, tandis que les autres sont classés avec des variabilités significatives. Avec davantage d'évaluations de modèles, jusqu'à 3200, ses performances sont améliorées pour classer correctement avec un faible degré de variabilité les 13 premiers paramètres responsables de 95 % de la variance totale, cependant, au-delà de ces paramètres, les performances sont médiocres. Au-delà de 3200, il semble tendre vers une meilleure précision et une estimation plus robuste avec plus d'évaluations de modèles.

Les performances de RBD-FAST sont conformes à ce que l'on trouve dans la littérature en ce sens qu'elles fonctionnent mieux sur les paramètres les plus influents. La méthode de Morris a un très bon potentiel non seulement pour les paramètres les plus importants mais aussi pour classer correctement les paramètres moins importants. Dans cette étude de cas, uniquement pour les trois premiers paramètres, la méthode de Morris a donné des résultats légèrement inférieurs à ceux de RBD-FAST ; cependant, au-delà de ces paramètres, elle a obtenu de meilleurs résultats en termes de robustesse et de précision. Ceci doit être confirmé sur d'autres études de cas. Globalement, la méthode de Morris est plus robuste et précise que RBD-FAST, et elle peut être utilisée avec moins de risques ; en particulier, elle fonctionne mieux que RBD-FAST en regroupant les paramètres les plus influents avec moins d'évaluations de modèles.

2.1 Introduction

Estimating a large number of model parameters through calibration could be unattainable due to identifiability problems caused by the interactions between these parameters. Considering a large number of parameters could also be cumbersome in terms of computational efficiency. Thus, a subset of the parameters is selected for calibration. Sensitivity analysis is a statistical method that aids in selecting this subset by estimating the influence of each parameter on the model outputs.

There are many different sensitivity methods that are available in literature. Some of these methods are accurate but computationally intensive such as Sobol method which is considered the most accurate one, others are computationally efficient but suffer from different drawbacks in terms of precision. In the context of building energy models, screening methods are extensively used due to their ability of retaining a sufficiently accurate ranking of the parameters with a relatively small number of model evaluations. There has been different developments in the field to improve the accuracy and computational efficiencies of the sensitivity methods. Thus it is important to quantify and evaluate how sufficient the accuracy of these more computational efficient methods is.

In this chapter, based on literature, different sensitivity methods are selected based on how frequently used in the field of building energy models and how promising they are. A detailed description of the different approaches retained is provided. The methods are applied to the case study of the concrete house located in Le Bourget-du-Lac, which is described in section 2.4 and a thorough comparison in terms of accuracy, robustness and computational efficiency is conducted.

2.2 Methods

Different existing sensitivity methods are briefly presented in chapter 1 and a general recommendation for the selection of appropriate methods is also provided.

Local sensitivity method was discarded since it does not account for parameters interactions, and it only allows for a limited perturbation around the nominal values of the parameters which consequently does not explore well the parameter space. In the following chapter, sensitivity methods that allow for better parameter space exploration are considered.

Screening methods account for the interactions between the parameters and explore better the parameters space with a reasonable number of model evaluations. Morris method is the most widely used in the context of building energy models due to its relative accuracy in screening the most important parameters and computational efficiency. Accordingly, Morris method is retained for further analysis in this chapter.

Global sensitivity analysis, and especially the variance based methods are known by their accuracy in the estimation of the sensitivity indices. RBD-FAST – a variance based method – overcomes the computational challenges of other methods. It has a considerable advantage compared to other methods like FAST or EFAST in this essence where it is able to estimate the main effects of the parameters with a data set of fixed *N* model evaluations (Gatelli et al. 2009). Under one set of simulations done on all the sampled points, it is sufficient to calculate the first indices for all the parameters. But one can pose the question about what the sufficient data set size *N* is to attain precise ranking of the parameters. Synthesised FAST is a method dedicated for a model with correlated parameters which is not the case with BEMs so it is discarded. Recently, some papers applied RBD-FAST to BEMs and it seems very promising (Goffart and Woloszyn 2021; Juricic 2020). Accordingly, and due to its computational advantage against other methods like FAST and EFAST, it is retained in this chapter for further analysis concerning its performance.

Among the variance based methods, Sobol method is selected as a reference method to which RBD-FAST and Morris' methods are compared since its results are the most accurate (Saltelli and Bolado 1998). This study allows to analyse the performance of both methods in terms of accuracy and stability through a robustness analysis. A detailed presentation of the Morris method is provided in appendix A. In the following section, a brief introduction to the variance-based methods is presented.

2.2.1 Variance-based methods

Sobol and Shukman (1993) used the Monte Carlo sampling method to solve the high dimensional integrals involved in calculating parameter indices. They started by decomposing the function f(x) (which is the model) into summands of increasing dimensionality as follows:

$$f(\theta) = f_0 + \sum_i f_i(\theta_i) + \sum_i \sum_j f_{ij}(\theta_i, \theta_j) + \cdots + f_{1\ldots k}(\theta_1, \ldots, \theta_k) \qquad (2.1)$$

where $\theta = (\theta_1, \theta_2, \ldots \theta_k)$ represents the k uncertain parameters. The model $f(\theta_1, \ldots \theta_k)$ is decomposed into different parts $f_i(\theta_i)$, $f_{ij}(\theta_i, \theta_j), \ldots, f_{1\ldots k}(\theta_1, \ldots, \theta_k)$, each is a part of the model output that is only affected by one indicated parameter i.e. $f_i(\theta_i)$ is the part of the model output $f(\theta)$ that is uniquely caused by the parameter θ_i, $f_{ij}(\theta_i, \theta_j)$ is the part of the model output $f(\theta)$ that is uniquely caused by the two parameter θ_i and θ_j. f_0 is a constant; it is the mean value of $f(\theta)$. The total variance of the model output $f(\theta)$ can be formulated as defined

$$D = \sum_i D_i(\theta_i) + \sum_i \sum_j D_{ij}(\theta_i, \theta_j) + \cdots + D_{1\ldots k}(\theta_1, \ldots, \theta_k) \tag{2.2}$$

It is shown how the total variance of the model can be decomposed to the partial variances caused by the parameters and their interactions. By dividing both sides of equation (2.2) by the total variance of the model output *D*, we obtain a summation of the indices of all the parameters (first summation) and their interaction (all other summations) and they all sum up to 1.

$$1 = \sum_i S_i(\theta_i) + \sum_i \sum_j S_{ij}(\theta_i, \theta_j) + \cdots + S_{1\ldots k}(\theta_1, \ldots, \theta_k) \tag{2.3}$$

S_i is the first order sensitivity index associated with the parameter θ_i which does not take into account the interaction between θ_i and all the other parameters. It is then deduced that S_i is equal to the variance of the output caused only by θ_i divided by the total variance of the model output.

$$S_i = \frac{D_i}{D} \tag{2.4}$$

S_{ij} is the second order sensitivity index associated with the interaction between the two parameters θ_i and θ_j. It measures the interaction effect between those two parameters on the variance of the model output (it is not the sum of the individual effects of θ_i and θ_j). The same applies to the interaction between three or more parameters $S_{1,2,..k}$. Any interaction index between any different parameters can be obtained by dividing the variance caused by this interaction on the model output $D_{i,j,..k}$ by the total variance *D*

$$S_{i,j,\ldots,k} = \frac{D_{i,j,\ldots,k}}{D} \tag{2.5}$$

From equation (2.3), if $\sum_{i=1}^{k} S_i$ approximately equals 1, this means that the parameters have few interactions as the second summation that correspond to interactions will approximately be zero.

The total sensitivity index is of a greater interest as it accounts for the variations caused by the parameter under question solely, in addition to its interaction with all the other parameters. Let us consider an example where there are two different parameters $X(x_1, x_2)$ in which the total variation *D* of the model output is calculated as the summation of the variances caused by x_1, x_2, and x_{12}. The total variance on the output caused by x_1 is follows:

$$D_{1Total} = D_1 + D_{12} = D - D_2 \tag{2.6}$$

Then, following the definition of the index, the total index effect of parameter x_i is the division of its total variance D_{1Total} by *D*, $\left(TS_{(1)} = 1 - \frac{D_1}{D}\right)$. This applies for all the parameters in the space as follows:

$$TS_{(\sim i)} = 1 - \frac{D_{(\sim i)}}{D} = 1 - S_{(\sim i)} \tag{2.7}$$

where $D(\sim i)$ is the sum of all the variances that did not include parameter *i* at all. The objective of this method is then to evaluate those two indices S_i and $TS_{(i)}$ for all the parameters in the space. To estimate these indices, a particular experimental design is generated. This design is contructed based on specific demarche in drawing the samples to account for first and second order indices. After this experimental design is constructed, different estimators can be used to estimate the Sobol indices. A detailed description on the experimental design and the different estimators is provided in appendix B.

2.2.1.1 RBD-FAST

Another method for computing the first order sensitivity indices is proposed (Tarantola et al. 2006). It is based on combining the random balance experimentations techniques elaborated by Satterthwaite (1959) with the FAST method. FAST is based on transforming the *k*-dimensional parameter space into one dimensional *s* space in the frequency domain using a transformation function G_i. Each parameter will be assigned an appropriate transformation function.

$$\theta_i = G_i(\sin f_i s) \tag{2.8}$$

where *i* goes from *1* to *k* (the number of uncertain parameters). *s* is the common variable for all the parameters and it oscillates from $(-\pi, \pi)$. In FAST, each parameter is assigned an appropriate angular frequency w_i. This can be thought of as a sampling technique, where each parameter is periodically oscillating at its corresponding frequency f_i. Assigning only one frequency to all the parameters will lead to generating sample points that are poorly distributed in the sample space, more specifically, in 2-D space, the sample points will only form a straight line. Tarantola et al. (2006) adapted the random balance design technique within RBD-FAST to perturb the generated samples. This allows to assign one single common frequency to all the parameters without generating a poor distribution. In another words, *N* points are sampled from the interval $(-\pi, +\pi)$, then the samples are perturbed randomly and differently for each parameter. Accordingly, the samples $S = \{s_1, s_2, \dots, s_j, \dots, s_N\}$ propagating in the search function to produce samples in the parameter space are differently ordered for each parameter

$$S^{(i)} = \left\{s_1^{(i)}, s_2^{(i)}, \dots, s_l^{(i)}, \dots, s_N^{(i)}\right\} \tag{2.9}$$

where *i* represents the parameter according to which the samples are being permuted, and *l* is the order of a permuted sample. Thus, through the search function, the values in the parameter space are correspondingly randomly permuted too. Eventually, the sample points will be well distributed in the sample space. Under these samples, the model is run *N* times.

$$Y(s_j) = f\left(\theta_1(s_{1j}), \theta_2(s_{2j}), \dots, \theta_k(s_{kj})\right) \quad \forall j = 1,2, \dots, N \tag{2.10}$$

To calculate the sensitivity index of a parameter θ_i, we apply an inverse permutation for all the parameters variables so that the θ_i values are reordered in an increasing form. This step orders θ_i in an increasing form but generates a new random permutation for all the other parameters. It is important to permute all the parameters corresponding to the same simulation at once and not only the parameter variable θ_i to ensure that the simulation results are consistent with the inputs. Correspondingly, the values of $Y(s_j)\ \forall\ j = 1,2, \dots, N$ are then reordered in the same manner as $\theta_i(s_{ij})$ and noted as $Y^R(s_j)$. The Fourier analysis and its coefficients are then estimated based on the reordered form of the model output based on the parameter in question as follows:

$$A_0 = \frac{1}{N}\sum_{j=1}^{N} Y^R(s_j) \tag{2.11}$$

$$A_k = \frac{2}{N}\sum_{j=1}^{N} Y^R(s_j)\cos(s_j k) \tag{2.12}$$

$$B_k = \frac{2}{N}\sum_{j=1}^{N} Y^R(s_j)\sin(s_j k) \tag{2.13}$$

The partial variance is then calculated using a predefined maximum harmonic *M*.

$$D = \sum_{l=1}^{M} F(w)_{f=l} = \sum_{l=1}^{M} F(l) \tag{2.14}$$

where $F(w)_{f=l}$ is the Fourier coefficient calculated in the previous step for the reordered samples according to parameter θ_i, and all these harmonics are summed up to obtain the variance just as in FAST.

$$D = 2\sum_{i=1}^{M}(A_i^2 + B_i^2) \tag{2.15}$$

For each parameter, the model output is reordered based on an increasing form and the Fourier coefficients are calculated to estimate the corresponding first order sensitivity index. It is cheaper computationally than the FAST and EFAST methods and the number of realisations is independent on the number of parameters due to the common frequency used (Gatelli et al. 2009). However, as mentioned earlier, this is true regarding the applicability of the method but needs to be investigated regarding the accuracy of the corresponding ranking. The parameters are sampled with the same frequency f. The maximum value allowed for the frequency in theory as stated by Tarantola et al. (2006) is:

$$f_{max} = (N-1)/2M \tag{2.16}$$

where N is the data set size. In this paper the frequency is set to 1 following the arbitrary choice of Tarantola et al. (2006).

This method results in a biased estimation of the indices S_i. This is due to the fact that each harmonic of *w* is related to the variance of all the other parameters and is falsely attributed to V_i. Tissot and Prieur (2012) provided a correction to this bias to account for the overestimation found in the original method:

$$S_i^c = S_i - \frac{\lambda}{1-\lambda}(1 - S_i) \tag{2.17}$$

where S_c is the corrected sensitivity index and, and $\lambda = \frac{2N_h}{N}$. As the sample size N and S_i increase the bias becomes less significant.

2.3 Methodology and criteria

To perform the comparison between the two selected methods (Morris and RBD-FAST), Sobol method is retained as a reference method since it is proven in literature that it is the most accurate variance-based sensitivity method. The methods are compared in terms of accuracy, robustness, and computation time.

It is important to mention that the quantity of interest, which is in our case the temperature profile, is divided into 24-hours time steps. That is that for each 24-hours period, the RMSE between the measurements and the simulations is taken. The sensitivity indices are then calculated for each parameter at each time step This generates a matrix of sensitivity indices; each column corresponds to a sensitivity vector of a separate parameter. To rank the parameters, the Euclidean norm of each vector is computed, and the parameters are ranked based on these values.

The ranking of Sobol method is retained as the reference ranking. The other methods are then analysed regarding how they can resemble the ranking obtained by Sobol method whether in obtaining the same exact ranking or in being able to at least cluster the most influential parameters even if they are not exactly ranked. This is done visually by plotting the paramerters as a function of their ranks. The Kandall rank correlation coefficient is also retained to quantify the similarities in the ranks between Sobol and the other two methods. The Kandall rank correlation τ is a statistic used to quantify the similarlity of the ranks between two data. The value of τ ranges from -1 to +1. A value of 1 means that the two data are ranked perfectly

similar, and a value of -1 means that the data are ranked in an opposite order. The Kandall correlation is given as follows:

$$\tau = \frac{n_c - n_d}{\frac{N(N-1)}{2}} \tag{2.18}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs. N is the number of data samples. Concordant pairs are two point (x_i, x_j) and (y_i, y_j) that satisfy $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$. The opposite is a discordant pair.

The Kandall correlation cannot be applied to a subset of the whole data set otherwise the results would not be reliable. For this, the Pearson correlation is also retained to be applied to subsets of the data. The Pearson correlation coefficient (PCC) is a measure of linear correlation between two set of data. A value close to unity means that the ranks between the two methods are close to be linearly correlated meaning that the tested method is able to resemble the ranks in the reference method. PCC is computed as follows:

$$PCC = \frac{k\sum_{i=1}^{k} r_{\theta_i} x_{\theta_i} - \sum_{i=1}^{k} r_{\theta_i} - \sum_{i=1}^{k} x_{\theta_i}}{\sqrt{k\sum_{i=1}^{k} r_{\theta_i}^2 - \left(\sum_{i=1}^{k} r_{\theta_i}\right)^2}\sqrt{k\sum_{i=1}^{k} x_{\theta_i}^2 - \left(\sum_{i=1}^{k} x_{\theta_i}\right)^2}} \tag{2.19}$$

where r_{θ_i} is the rank of parameter θ_i in the reference method, and x_{θ_i} is its rank in either Morris' method or RBD-FAST, and k is the total number of parameters. The Pearson correlation is applied to all the parameters sequentially starting from the first two parameters $k = 2$, and increasing *k* until at all the parameters are included. This enables to asses the performance of each method with different clusters of parameters: the methods performance on the most influential parameters only and on all the parameters.

The robustness of each method is also analysed. The stability in ranking the parameters after multiple repetitions is essential to ensure that the selection of the parameters is reliable. To assess this behaviour, each method is repeated several times with increasing data set sizes (number of model evaluations). Note that for Morris' method, it is important to select the number of levels and the repetitions beforehand which results in a specific data set size required to execute the method. However, in contrary to Morris and Sobol methods, RBD-FAST does not depend on a specific data set size to be executed, which makes it easy to choose a size that is consistent with that required by Morris' method. All of these runs form a data set comprising

the rank of each parameter after each run. It is useful to compute the standard deviation of the ranks of each parameter to analyse the stability of the methods.

The methods robustness and accuracy enables to have a better idea about the required data set sizes needed by each method to attain reliable results. This in turn reflects on the computational efficiency of each method. The time accounted for is the simulation time of the BEM to generate the required data set. The time taken by the method itself to compute the sensitivity indices is not taken into account since it is negligible.

2.4 Case study

The studied building corresponds to the I-BB house (Concrete construction) of the INES (National Institute of Solar Energy) "INCAS" platform, located in Le Bourget-du-Lac (France) (Figure 2.1). The net floor area is 89 m² on two levels. The house is designed to match the performance of the "PassivHaus" label, thanks to high insulation, very low thermal bridges, and high-performance glazing.

Figure 2.1: Sketch of the IBB house

The interior dimensions are 7.5 m in length and 6.5 m in width. The house is built on a crawl space 80 cm high and is surmounted by an unheated attic. Its orientation is offset 15 ° counter clockwise from the north-south axis. The roof is two-sided with a North / South orientation and an overhang of 60 cm to the east, west and north. The south facade of the building includes a large glazed surface (28 % of the facade) protected from the sun by a balcony with a width of 1.3 m and a roof overhang of 1 m. The east and west facades have a glass area ratio of 5 % and 10 % respectively.

The I-BB house is highly insulated (external insulation in walls). 20 cm of extruded polystyrene are added to the external walls of the house that are composed of 15 cm of shuttered concrete. The ground floor consists of 20 cm extruded polystyrene, 16 cm cast-in place concrete, and 8 cm of concrete screed. For more details about the compositions of the different parts of the building, see Spitz (2012).

The house is equipped with a double-flow type CMV (Controlled Mechanical Ventilation) with heat recovery and two fans: one for the fresh air circuit and the other for the stale air circuit. Fresh air is blown into the living room and bedrooms, while vents in the toilets and bathroom ensure the extraction of stale air. The plate heat exchanger allows up to 90 % heat recovery. A by-pass allows cooling by direct ventilation. Heating is provided by an electric resistance located at the start of the fresh air distribution network.

An experimental campaign has been conducted to measure the temperature profile in the building. The considered period (January 1st to April 22nd, 2012) was subdivided into six scenarios (Table 2.1), with different physical phenomena. Parameters, such as the heating setpoint, the opening / closing of the shutters, or the mechanical ventilation flowrate were modified. The ventilation mode was also varied between a by-pass (BP) mode and single-flow ventilation. The ventilation flowrate when turned on also changed in the last scenario where rate-a corresponds to 110 m^3/hr, and rate-b corresponds to 150 m^3/hr.

Table 2.1: Illustration of the different scenarios of the case study

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
T° setpoint	20°C	Off	Off	24°C	Off	Off
Ventilation mode	BP	Off	Off	BP	BP	BP
Ventilation flow rate	Rate-a	0	0	Rate-a	Rate-a	Rate-b
Shutters	Closed	Closed	Closed	Closed	Open	Open
	01/01/2012	16/01/2012	09/02/2012	20/02/2012	15/03/2012	31/03/2012	22/04/2012

The sensitivity analysis is applied to select the parameters on which calibration will then be performed. Thus, it is important to peform the sensitivity analysis on the scenarios that will be considered for training the calibration model. As in Robillart (2015), scenarios 1, 2, 3, and 5 are considered training scenarios on which calibration process will be performed. Scenarios 4 and 6 are chosen as validation data because they are respectively close to the 1st and 5th scenarios. We can thus evaluate the behaviour of the calibrated model under relatively

similar experimental conditions. Accordingly, in this chapter the sensitivity analysis will only be applied on the training scenarios.

2.5 Results and discussion

The bounds assigned to the parameters are taken from Munaretto (2014) and are presented in appendix. Some parameters that Munaretto included are not considered in this study such as the setpoint temperature: only those that can be calibrated are included. The same bounds are set for Morris methods, however, for Sobol method and RBD-FAST, normal distributions over these bounds are considered. The standard deviation of each parameter is selected so that the bounds of Munaretto (2014) are within 2 standard devations from the mean. This ensures the consistency in the comparison between the methods. It is preferred to use normal distributions instead of uniform for the variance based methods to account for their capability of incorporating such distributions.

Sobol method is executed with 4000 samples following Pannier (2017). She assessed the convergence of the total sensitivity indices with different sample sizes starting from 100 ending with 5000. The study showed that the total indices started converging from a sample size of 2000. She used in her study a sample size of 5000 to ensure the reliability the ranking. In our study, the number of parameters included is larger than those used in her study, so, a smaller sample size of 4000 is used to reduce the computation time while ensuring reliable ranking . There are 113 parameters in the model, which means that 460,000 simulations are executed.

Figure 2.2 shows the parameters temporal ranking obtained via Sobol method. It only presents the parameters that explain 90 % of the total variance: ventilation flowrate ($\dot{V}$), heating power (Q_p), specific heat of wall concrete $(c_{p,concW})$, conductivity of polystyrene wallmate (λ_{polW}), dissipated heat (Q_d), solar albedo[3] (Alb), specific heat of concrete screed $(c_{p,concS})$, and conductivity of polystyrene Styrofoam (λ_{polS}). The importance in the four training scnearios is shown here. The last training scenario is combined in the plot with the first three eventhough in reality they are separated by one testing scenario.This explains the sudden increase in the importance at day 50. The first scenario corresponds to the first 16 days, then

[3] Reflectivity of the ground around the building

the next 34 days correspond to the second and third scenarios, and the rest correspond to the fifth scenario.

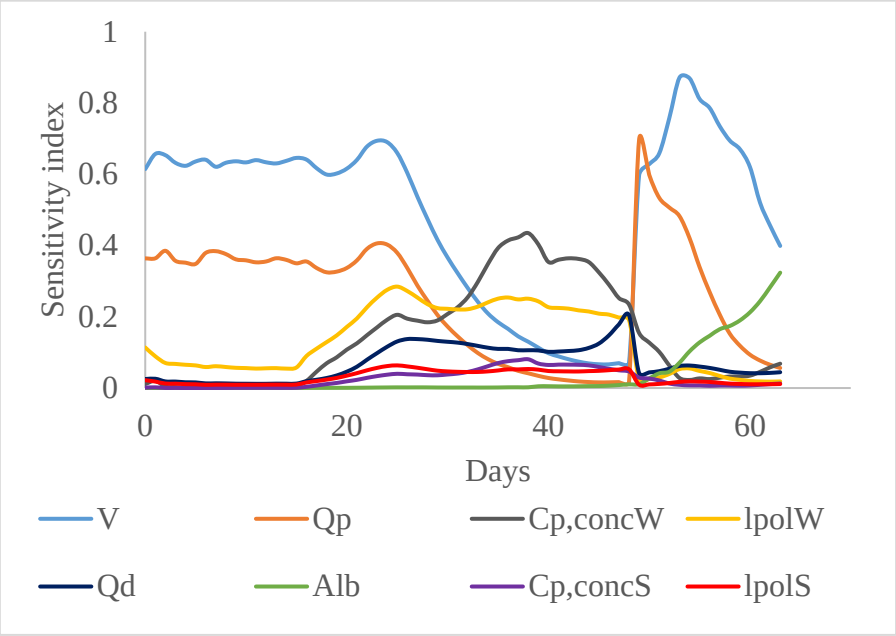

Figure 2.2: Temporal sensitivity indices

This figure allows to visualise the importance of each parameter and how it varies with time. On the y-axis is the total sensitivity index computed at each 24-hours time step. For the first scenario: first 16 days, it is obviously observed that the ventilation flow-rate and the heating power are the most important parameters compared to the rest. The building model considered being single-zone, scenarios 2 and 3 are identical. Therefore, the results are similar for these two scenarios. It can be observed (day 16 to 50) that the importance of both the ventilation flow-rate and the heating power decreases gradually since in these two scenarios the heating power is turned off and there is no ventilation. The building is kept in free evolution during these two scenarios. Accordingly, the importance of the insulation conductivity and the specific heat of the concrete becomes more obvious in these scenarios. It is also important to mention, that during all the experimental setup, the dissipated heat (representing internal gains) is constant and is not changed, however, its importance is only significant during the free evolution of the building.

At the end of scenario 3, a sudden change in the parameters importance is observed. That is because, the sensitivity indices are computed only for scenarios 1, 2, 3, and 5: the variation in the sensitivity indices caused by scenario 4 are not shown. The influence of the heating power at the beginning of scenario 5 is high, and then directly starts decreasing. The reason behind

that, is that the heating power is turned on in scenario 4 and then turned off at the beginning of scenario 5. The ventilation existed in both scenario 4 and scenario 5 which explains its high importance at scenario 5 where it does not decrease as is the case for the heating power. However, its importance is not constant as it is the case in the first scenario. The reason behind that is that the solar albedo started sharing some importance with the ventilation starting from scenario 5. This is due to the fact that the shutters in scenario 5 are open.

After computing the length of each vector using the Euclidean distance, the parameters are ranked from the parameter having the largest length to the parameter having the smaller length. Figure 2.3 shows the ranking of the parameters responsible for 90 % of the total variance.

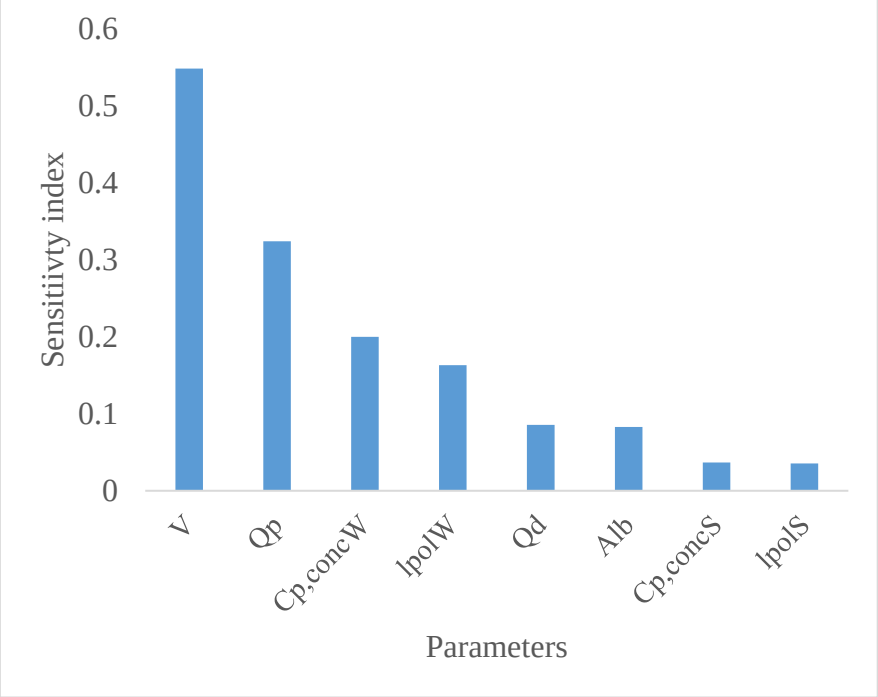

Figure 2.3: Parameters ranking with Sobol

In a first step, the aim is to assess the performance of Morris method and RBD-FAST with a sufficient sample size and in a second step to assess their robustness with different sizes. Accordingly, Morris' method is launched with 60 repetitions which corresponds to 6840 model evaluations. This is close to the number of repetitions (50) sufficient for a precise ranking of the parameters as found by Pannier (2017). To undergo a reliable comparison, the data set generated for RBD-FAST is of the same size.

The first interesting indicator is to count the number of parameters that are ranked exactly as Sobol method. With Morris' method, 33 parameters had similar ranks as estimated by Sobol method; with RBD-FAST 14 parameters are correctly ranked. Figure 2.4 and Figure 2.5 show a scatter plot for the parameters' ranks of Morris' method and RBD-FAST versus Sobol method respectively.

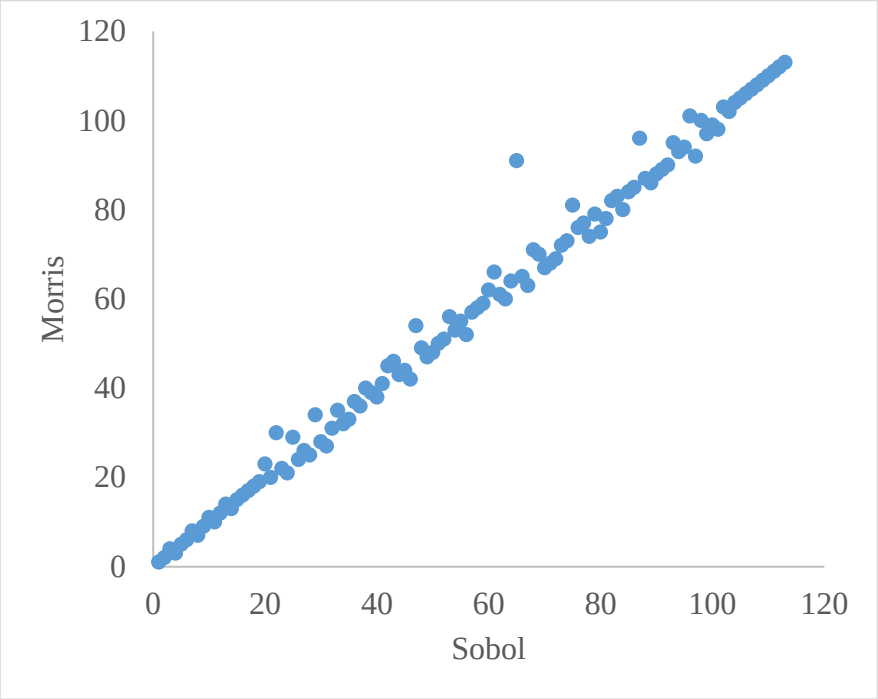

Figure 2.4: Correlation between Sobol and Morris ranking

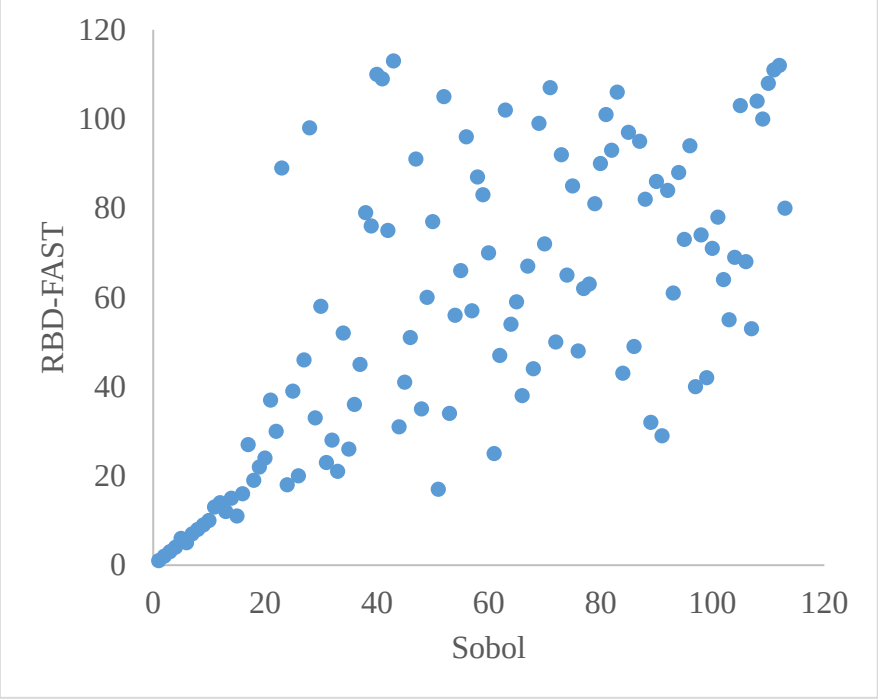

Figure 2.5: Correlation between Sobol and RBD-FAST ranking

It is clearly depicted how Morris' method fits better to Sobol method than RBD-FAST, even with the parameters that are not exactly ranked as Sobol method, they do not diverge significantly as with RBD-FAST, where a random dispersion exists. The degree of similarity in the ranks between Morris' method and Sobol method is estimated via Kandall rank coefficient and it yielded a value of 0.95. On the contrary, the coefficient between RBD-FAST and Sobol method is 0.5. The difference between both methods is quite large and it can be clearly visualised in the scatter plots. It is also clearly depicted that for the most influential parameters, the relation between Sobol and RBD-FAST seems to follow a linear behaviour. This points out the importance of analysing not only the ranks of the whole parameter space, but also to focus the comparison on ranking the sets of the most influential parameters.

To quantify the differences between both methods with different set of parameters, the Pearson correlation coefficient (PCC) between Sobol and these methods is computed sequentially starting from ranking the three most important to ranking all the parameters. The reason behind starting with three parameters and not from two is that a linear relationship could always be found for two data points.

Figure 2.6 shows how the correlation between Sobol and the other methods change with increasing number of parameters. On the one hand, it shows a close linear correlation between the ranks of the reference method and RBD-FAST with a limited number of parameters (around 20); then, the coefficient of linearity decreases with increasing number of parameters to reach a value of 0.68. The evolution of PCC is not perfectly smooth: PCC tends to decrease with increasing number of parameters but there exist increases and decreases in between. This could be related to the randomness in ranking the less influential parameters compared to the reference method. On the other hand, PCC between Morris' method and the reference method stays approximately constant around 0.98 with increasing number of parameters. At the beginning, there exists a slight decrease to 0.8 and then the PCC increases again. The reason behind this decrease is that, there is a swap in ranking the third and the fourth parameter (specific heat of concrete, and conductivity of polystyrene) with Morris' method, which is translated by a decrease in the value of PCC when computed on the first three and four parameters. Since Morris' method functions well in resembling the ranks of Sobol method with very small variation, the PCC tended to increase towards a value very close to unity. The intersection between the two curves in Figure 2.6. occurs at the fourth parameter which explains this analysis.

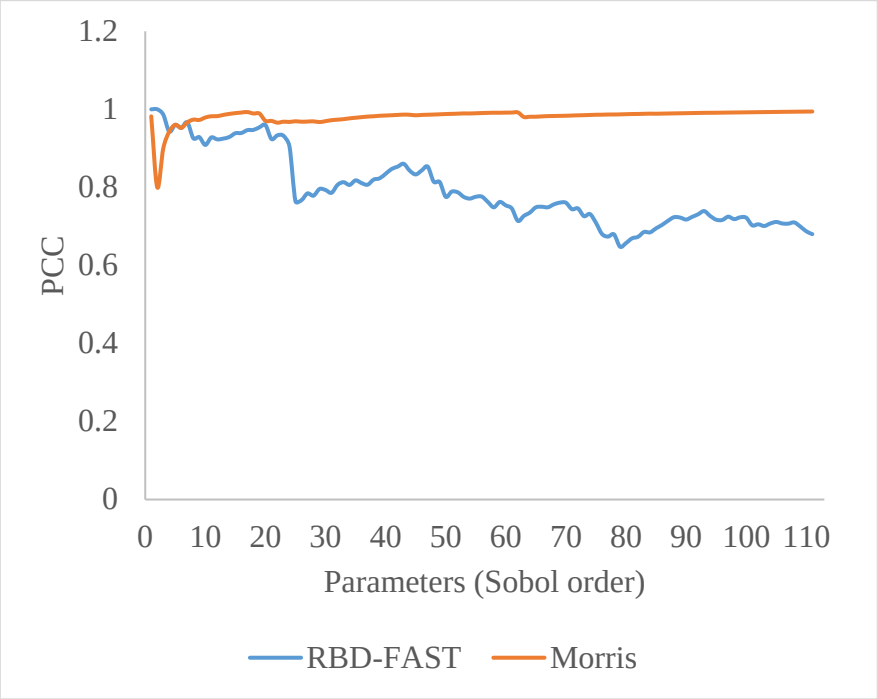

Figure 2.6: PCC correlation between Sobol, Morris, and RBD-FAST rankings

The sensitivity indices and the resulting rankings could vary if the method is repeated with a different data set. Thus, it is also important to analyse how accurate and robust each method is against repetitions and how the performance changes with different sample sizes. For this purpose, Morris' method is performed on the same case study with different repetition starting from 5 ending with 60. Correspondingly, RBD-FAST is repeated with different sample sizes equivelent to those indicated with Morris' method repetitions. The robustness of each method is evaluated by estimating the variabilities in the parameters rankings.

Figure 2.7 shows the Kandall τ rank coefficient for each execution of the two sensitivity methods. Morris' method shows a constant performance with an increasing number of repetitions, where the value of τ remained almost constant around 0.98. This also indicates a good robsutness against number of repetitions with negligible variabilities. In the graph, the number of model evaluations correspoding to Morris' repetitions is displayed and not the number of repetitions. On the other side, with RBD-FAST the value of τ shows a significant variability compared to Morris method with an increasing trend. This means that with a larger data set, RBD-FAST tends to rank the parameter more accurately however, with a significant degree of variability: the standard deviation for the values of τ for RBD-FAST is 0.078, compared to 0.005 for Morris' method.

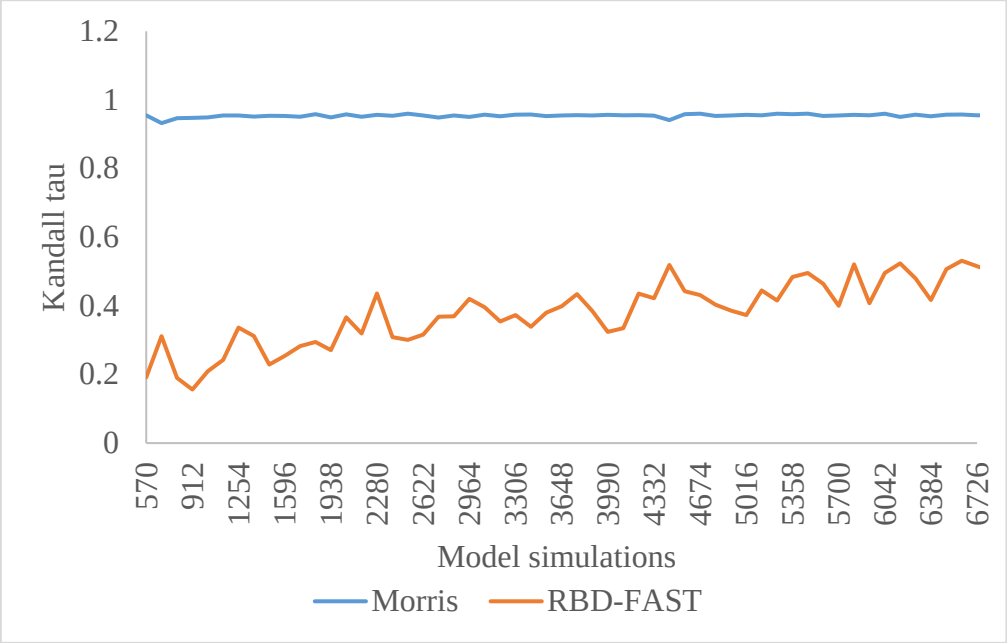

Figure 2.7: Kandall τ correlation for Morris and RBD-FAST

Computing the coefficient on all the parameters gives some sort of information on how each method accurately ranks the ensemble of all the parameters. Same as previously done, the PCC is computed sequentially for each execution of the methods. The aim is to analyse the robustness of each method against number of simulations in ranking the most influential parameters. To this end, it is interesting to assess the accuracy of both methods to ranking the parameters responsible for 90 % (first 8 parameters) and 95 % (first 13 parameters) of the total variance estimated by Sobol method. Figure 2.8 shows the PCC computed for both methods on the first eight parameters. It shows an approximate similar performance for both methods in ranking the first 8 parameters with a slight advantage to RBD-FAST. The variations in the PCC for Morris' method is due to the variations in ranking $c_{p,concW}$, and λ_{polW} as shown in Figure 2.9.

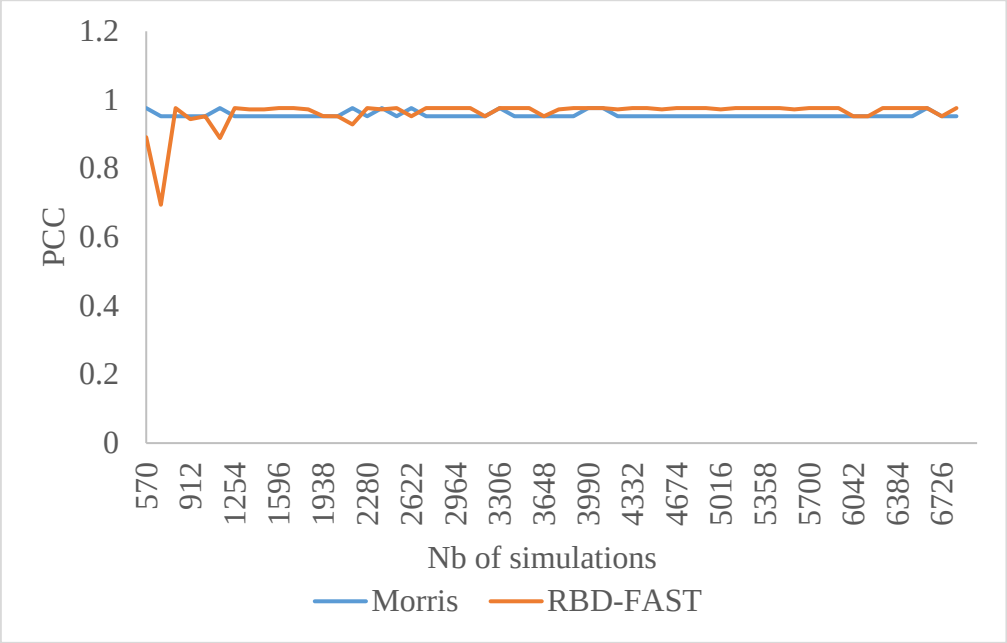

Figure 2.8: PCC of Morris and RBD-FAST on first 8 parameters vs increasing simulations

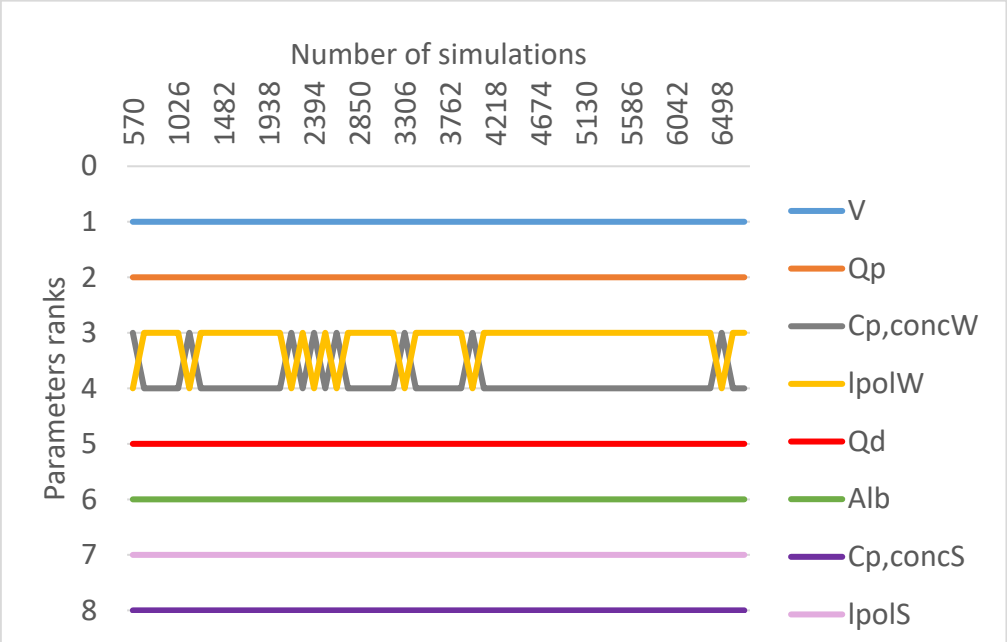

Figure 2.9: Morris ranking of the first 8 parameters

It could be visualised that the PCC of Morris' method in Figure 2.8 increases only when these two parameters are ranked as with Sobol method. The last two parameters are consistently ranked, however different to the order estimated by Sobol method. RBD-FAST is more consistant at ranking the first 4 parameters as shown in Figure 2.10. However for the last two parameters, there exists significant variabilities in the ranks. The dissipated heat and the albedo are consistently ranked, however differ from the order estimated by Sobol method. To summarise, in both methods, four parameters are precisely ranked in a consistent manner and

the other four are not. Given these observations, and the results of PCC shown in Figure 2.8, it could be said that RBD-FAST performed in this case slightly better than Morris' method in ranking the first 8 parameters. One more thing to add is that RBD-FAST at some runs is not able to cluster the conductivity of polystyrene as one of the eight most influential parameters estimated by Sobol method which is not the case with Morris' method. On the contrary, the first three parameters are consistently ranked by RBD-FAST and not by Morris' method.

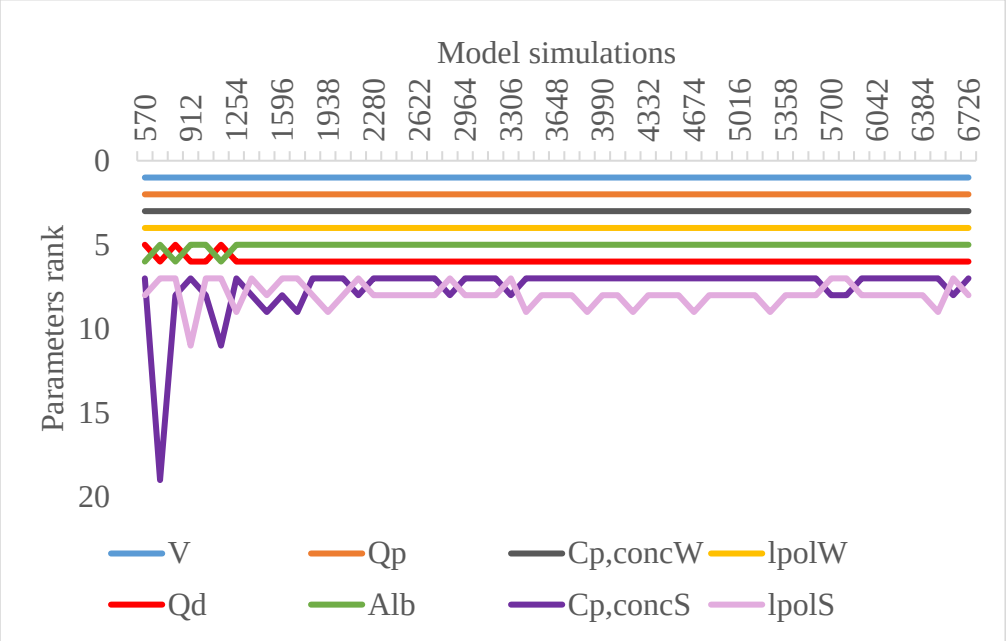

Figure 2.10: RBD-FAST ranking of the first 8 parameters

Morris' method shows a consistent rank for the first 13 most influential parameters with different runs. Figure 2.11 shows the rank of the 9th to the 13th most important parameter: concrete wal thickness (t_{concW}), thermal bridge (ψ), window heat transfer coefficient (U_w), thickness of polystyrene Wallmate (t_{polW}), and the specific heat of reinforced concrete ($c_{p,concR}$). It is clearly depicted that those parameters are consistently clustered as the most influential ones. On the contrary, RBD-FAST does not do as well with the group responsible for 95% with smaller sample sizes. To quantify this better, the PCC of Morris' method and RBD-FAST in ranking the parameters responsible for 95 % of the total variance is depicted in Figure 2.12. Morris performs better and more consistently with small sample sizes: less than 3200. The performance of RBD-FAST is improved with bigger sample size: PCC approaches that of Morris' method. However, unlike Morris' method, there exists a greater degree of variability on the PCC value even after a sample size of 3200: the standard deviation of PCC

estimated with a sample size of 3200 and above for RBD-FAST is 0.027 compared to 0.0036 for Morris' method.

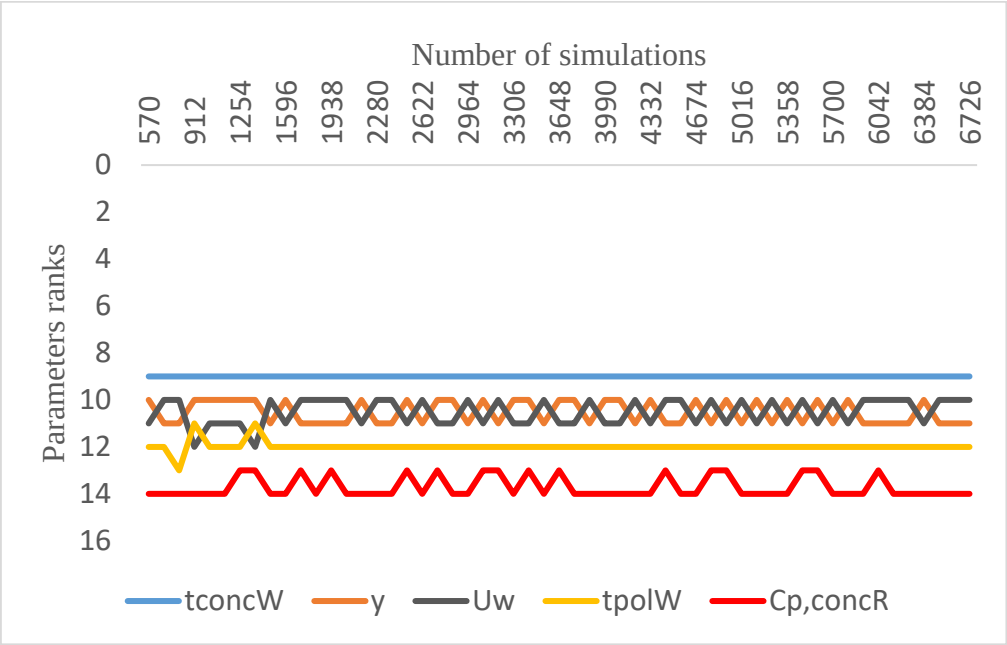

Figure 2.11: Morris ranking of the parameters contributing to 95 % of the total variance

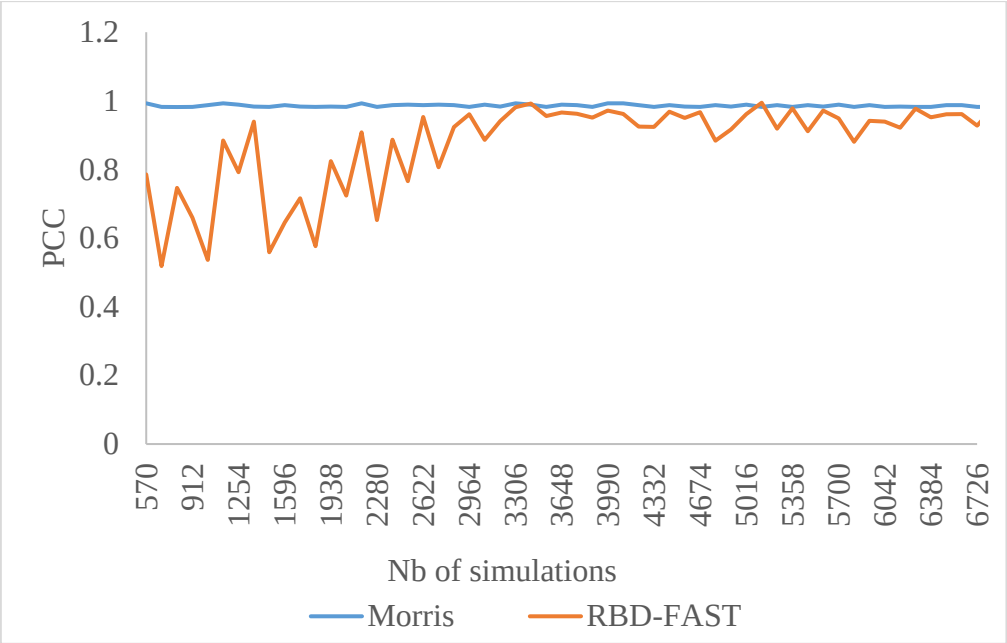

Figure 2.12: PCC of Morris and RBD-FAST on the first 13 parameters vs increasing simulations

One more important observation is that unlike Morris' method, RBD-FAST is not capable of consistently clustering the 13 most influential parameters as depicted in Figure 2.13 This gives an advantage for Morris' method against RBD-FAST.

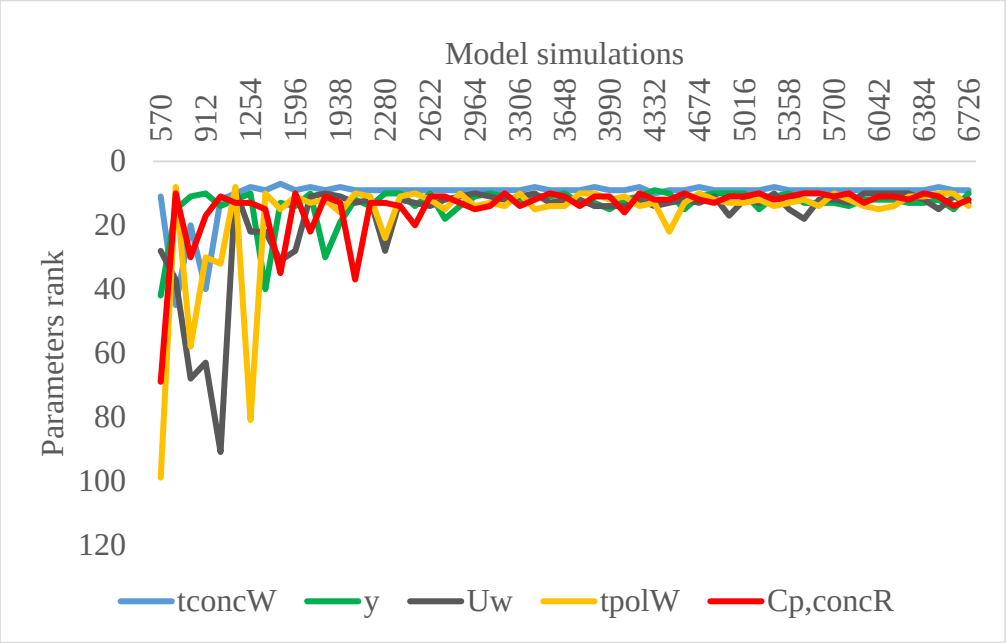

Figure 2.13: RBD-FAST ranking of the parameters contributing to 95 % of total variance

According to these observations, it could be said, if only a very small set of parameters is needed (the first two or the first three), then RBD-FAST could be a better choice, however if more than that is needed, then Morris' method could be a better choice given that it shows a better performance in clustering the important parameters.

2.6 Conclusion

The objective of this chapter is to analyse the performance of efficient sensitivity analysis methods compared to Sobol method. Morris method is a screening method that is widely used especially in the field of building energy models. RBD-FAST has been recently applied in the field and it is computationally efficient compared to other variance-based methods. A comparison is executed between these two methods in terms of accuracy, computational efficient and robustness. Sobol method is considered as the reference method.

Morris' method shows a very good performance even with a small number of repetitions yielding approximately similar ranks as Sobol's method even on the relatively non influential parameters. It is very good at clustering the most influential parameters even if at some points it does not exactly rank all the influential parameters that are ranked by Sobol's method. Some variation in the ranks of some influential parameters is observed but is negligible. However, the third and the fourth parameters are swapped very often with different runs which is not the case with RBD-FAST.

RBD-FAST shows a good performance in ranking the parameters responsible for 90 % of the total variance estimated by Sobol's method, especially for the first three parameters. These parameters are consistently clustered with a relatively small number of model evaluations compared to Sobol's method while the rest are ranked with significant variabilities. With more model evaluations, up to 3200, its performance is enhanced to correctly rank with a small degree of variability the first 13 parameters responsible for 95 % of the total variance, however, beyond these parameters the performance is poor. Beyond 3200, it seems to tend towards better accuracy and more robust estimation with more model evaluations.

RBD-FAST performance is consistent with what is found in the literature in that it performs better on the most influential parameters. Morris' method has a very good potential not only for the most important parameters but also in correctly ranking less important parameters. In this case study, only for the first three parameters, Morris' method performed slightly worse than RBD-FAST; however, beyond these parameters, it performed better in terms of robustness and accuracy. This needs to be confirmed on other case studies. Globally, Morris' method is more robust and accurate than RBD-FAST, and it can be used with less risk; especially, it performs better than RBD-FAST in clustering the most influential parameters with less model evaluations.

Chapter 3

Calibration methods assessment

This chapter focuses on various Bayesian calibration methods existing in literature. A detailed explanation on the statistical background of these methods is provided. Then, they are applied on a virtual case study and compared in terms of precision and computational efficiency. Their performance in identifying the true values and in enhancing the calibrated model precision is studied.

Résumé du chapitre

Le calibrage bayésien a suscité l'intérêt de nombreux chercheurs dans le domaine de l'énergétique des bâtiments. Il permet de connaître l'incertitude dans la prédiction d'un modèle à partir des distributions a posteriori qu'il fournit sur les paramètres d'entrée. Cela permet d'estimer le niveau de confiance dans les prédictions du modèle utilisé. Il s'agit d'une méthodologie appropriée dans le cadre de la rénovation des bâtiments et des contrats de performance énergétique.

Cinq méthodes de calibrage bayésienne (ABC-PMC, APMC, ABC-RF, CATMIP, Adams) sont sélectionnées dans la littérature, en fonction de leur popularité et de leur capacité à être parallélisées. Une analyse comparative entre ces algorithmes est effectuée en termes de précision et d'efficacité de calcul. La même étude de cas utilisée dans le chapitre précédent est conservée. En conséquence, les résultats de l'analyse de sensibilité sont pris en compte pour sélectionner les paramètres de calibrage. Les données virtuelles sont utilisées pour effectuer le calibrage. Cela permet d'analyser les performances des méthodes dans des conditions contrôlées sans l'effet d'erreurs et d'incertitudes de mesure et de modèle supplémentaires, et également d'évaluer les performances des méthodes dans l'identification des vraies valeurs des paramètres (qui sont connues car elles constituent les entrées de la simulation ayant produit les mesures virtuelles).

Les critères utilisés pour mener cette comparaison sont la distance euclidienne pondérée entre les distributions a posteriori des paramètres et leurs vraies valeurs, et la RMSE moyenne de la propagation a posteriori. Ces critères sont appliqués sur les postérieurs obtenus par chaque méthode. De plus, ils sont également appliqués sur les distributions de paramètres obtenues dans les itérations intermédiaires. Cela permet d'analyser le comportement des méthodes avec un nombre croissant d'évaluations de modèles.

Les résultats ont montré que APMC et Adams surpassent les trois autres algorithmes. La raison de cette performance différente est liée à l'échantillonneur adapté et non à la dépendance à la fonction de vraisemblance. ABC-RF a montré une meilleure performance avec un nombre inférieur d'évaluations de modèles mais la pire performance avec un nombre croissant de simulations. Mais dans cette évaluation, le nombre d'arbres et la taille des feuilles n'ont pas été augmentés en fonction du nombre de simulations : les valeurs par défaut du module fourni dans

l’environnement R ont été considérées. Par conséquent, une enquête plus approfondie sur ABC-RF serait utile compte tenu de ses bonnes performances relatives avec moins de simulations.

3.1 Introduction

Bayesian calibration applied to building energy models have gained the interest of many researchers in the field. It naturally accounts for the uncertainty in the model prediction through the posterior distributions that it provides. This allows estimating the level of confidence in the predictions of the BEM used. It is an appropriate methodology to apply in the context of building renovation and energy performance contracting.

Using Bayesian calibration, the parameters are modelled with a distribution that explains the belief about their values called prior distributions. These parameters are then estimated by updating these distributions. Accordingly, the estimated parameters take the form of distributions called posteriors. As described earlier, an analytical form of the posterior is often impossible to attain. Due to this, the posterior could be estimated by drawing samples from it. Then if these samples are found to be normally distributed for example, their mean and standard deviation can be taken to identify a Gaussian posterior law. Different types of samplers exist and have been applied in the Bayesian context. The choice of the sampler has a considerable effect on the accuracy and computational efficiency of calibration.

Different calibration methods are selected from literature, according to their popularity and their ability to be parallelised. A comparative analysis between those algorithms is done in terms of accuracy and computational efficiency. To perform this analysis, the same case study on which the sensitivity analysis is performed is retained and the results of the sensitivity analysis are taken as the basis for the calibration. The comparison in this chapter is executed on virtual data. This allows to analyse the performance of the methods in controlled conditions without the effect of additional measurement and model errors and uncertainties.

Firstly, the methods retained and implemented in this chapter are illustrated with sufficient details in section 3.2. Then the methodology and comparison criteria on which the conclusions are drawn is presented in section 3.3. Then the results are presented and discussed in section 3.4.

3.2 Methods

Bayesian calibration is divided into two families: likelihood-dependent and likelihood-free approaches also called approximate Bayesian computation (ABC). Recently, many

researchers have been focusing on improving the ABC methods. Various updated more robust versions has been published some of which are very robust in terms of hyper parameters. Still in the context of building energy efficiency, there is no sufficient applications of these methods and both families of Bayesian calibration are not compared.

In this chapter, five algorithms from both families with different sampling techniques are selected, according to their popularity and their ability to be parallelised. A comparative analysis between those algorithms is done in terms of accuracy and computational efficiency.

3.2.1 Likelihood dependant approaches

Recently, with the increase of the application of methods that require sampling from an unknown distribution, many studies have focused on developing better sampling techniques. This section walks through the different types of samplers existing with detailed description of those that are used in the thesis.

One class on which many subsequent samplers are based is the Marchov chain Monte Carlo (MCMC). One of the well known implementations of this sampler is the Metropolis-Hastings technique. This approach describes a sequence of possible events in which the probability of each event depends only on the state attained after the previous event. The sequence of events here means the sequence of samples generated. To understand the concept of this sampler, it is important to define two main aspects. The first is the proposal distribution, also called transitional kernel $K(.)$. The objective of this distribution is to generate new proposal at each step given the sample at the current step. The proposal here means a new value for the parameter to be calibrated. The second is the Metropolis-Hastings acceptance-rejection ratio α which is used to decide whether to accept or reject the proposed parameter value as a sample from the posterior:

$$\alpha = \frac{p(\theta^*|Z)\, K\left(\theta^{(i-1)}\middle|\theta^*\right)}{p\left(\theta^{(i-1)}\middle|Z\right) K\left(\theta^*\middle|\theta^{(i-1)}\right)} \tag{3.1}$$

θ^{i-1} is the sample at the previous step and θ^* is the sample to be accepted or rejected. If a symmetric kernel is chosen, then the ratio $K\left(\theta^{(i-1)}\middle|\theta^*\right)/K\left(\theta^*\middle|\theta^{(i-1)}\right)$ will cancel to 1. Normally, a Gaussian distribution is used as a transitional kernel having a mean $\theta^{(i-1)}$ and a

variance. The variance is one of the hyper-parameters of the sampler based on which the global performance of the MCMC can change.

As shown in algorithm-1, to launch the sampler, the first sample $\theta^{(i-1)}$ is drawn from the prior distribution and its probability being a sample from the posterior is estimated from equation (1.5) from chapter 1. This sample serves as the starting point of the Marchov chain. The transitional kernel in then used to draw another sample $\left(\theta^{*}\middle|\theta^{(i-1)}\right)$. The generated sample is called proposal. Equation (3.1) is then computed and a random number u is generated from a uniform distribution over the interval [0,1]. If α is greater than the random value, then the proposed sample is accepted and the chain continues, otherwise the chain stays at the current sample and generates a new proposal θ^{*} through the transitional kernel. In other words, the proposal that is more probable than the current sample will always be accepted. However, if the proposal is less probable than the current sample, sometimes the proposal will be accepted and sometimes it will be rejected; the larger the relative drop in probability, the more likely the proposal will be rejected. This methodology will then keep the sampler most of the time around the high density regions of the posterior.

In the computation of α, the normalising constants present in equation (1.4) in chapter 1 cancel out which avoids the trouble of computing it. The Markov chain allows the sampling from the prior PDF only the first iteration and then all the other samples are drawn from the kernel that depend on the previously accepted sample which makes them closer to the posterior than to the prior PDF. Understanding the concept behind algorithm 3.1 is necessary to understand the following more developed algorithms.

Algorithm 3.1

1. Initialise by proposing a sample θ^0 from the prior $p(\theta)$ and compute $p(\theta|Z)$
2. Identify the length of the random walk N
3. **for** $i = 1, \dots, N$:
    a. Generate a new proposal θ' from $K\left(\theta^* \middle| \theta^{(i-1)}\right)$
    b. Compute $\alpha = \frac{p(\theta^*|Z)\, K\left(\theta^{(i-1)} \middle| \theta^*\right)}{p\left(\theta^{(i-1)} \middle| Z\right) K\left(\theta^* \middle| \theta^{(i-1)}\right)}$
    c. Generate a random number u from a uniform distribution $U(0,1)$
    d. **if** $\alpha \geq u$:
        i. Set $\theta^i = \theta'$
    e. **else**:
        i. Set $\theta^i = \theta^{i-1}$

An efficient MCMC sampler NUTS proposed by Hoffman and Gelman (2011) is recently used for building energy models. NUTS can converge with a relatively small number of samples compared to Metropolis-Hastings. However, the number of samples generated is not the total number of likelihood computations. In order to attain the posterior samples, NUTS requires numerous computations of the likelihood gradients, which is computationally inefficient given that it is un-parallelisable and that in this chapter, no metamodel is used to replace the original one. For this reason, sequential Monte Carlo samplers (SMC) are preferred since they benefit from their capability of being easily parallelised.

MCMC can be extended to be applied sequentially. Ching and Chen (2007) proposed the transitional MCMC (TMCMC) sampler. The general concept is that instead of directly sampling from the posterior, intermediate distributions could firstly be sampled before the posterior is reached. This is called the sequential Monte Carlo (SMC) or particle filter. Instead of starting with one draw from the prior and constructing one chain, *N* samples (also called particles) corresponding to *N* chains are drawn and assigned weights; each sample is moved in the parameter space to explore it using an appropriate transition kernel and then accepted or rejected following the Metropolis Hastings ratio. However, instead of directly being accepted or rejected as a sample from the posterior, at each iteration, the samples are validated as to follow different distributions. At each iteration the distribution to which the samples belong will be closer to the posterior and further from the prior. Those distributions are controlled with an annealing

parameter β that indicates how far or close each distribution is to the previous. These distributions are determined as follows:

$$f_t(\theta|Z) \propto p(\theta)p(Z|\theta)^{\beta_t} \qquad \beta_t = 0, \dots, 1 \tag{3.2}$$

where t is the iteration indicator. If $\beta_t = 0$, then the samples generated estimates the prior distribution, and if $\beta_t = 1$, then the samples are generated from the target posterior distribution. The sequence of β_t that defines the intermediate distributions has a huge effect on the performance of the sampler. On the one hand, if there exists a great difference between two subsequent distributions, then the sampler might collapse due to degeneracy problem; all the particles collapse to one duplicated particle. On the other hand, if the subsequent distributions are very close to one another, then too many intermediate distributions are defined before reaching the target distribution which makes it computationally inefficient.

For this purpose, Ching and Chen (2007) proposed to define the value of β adaptively at each step such that the coefficient of variation (cov) of the particles weights is unity.

$$cov\left\{w(\theta_{t,i}) : i = 1, \dots, N\right\} = 1 \tag{3.3}$$

The weights $w(\theta_{t,i})$ are computed by taking the ratio of the probability of each particles at the next distribution β_{m+1} to their probability at the current iteration β_m.

$$w(\theta_{t,i}) = \frac{p(\theta_{t,i})\, p(Z|\theta_{t,i})^{\beta_{t+1}}}{p(\theta_{t,i})\, p(Z|\theta_{t,i})^{\beta_t}} = p(Z|\theta_{t,i})^{\beta_{t+1}-\beta_t} \tag{3.4}$$

The sampler proceeds from a current iteration t to a subsequent iteration $t+1$ through resampling. Resampling here means to choose particles from iteration t and move them to iteration $t+1$. This selection is based on the weights $w(\theta_{t,i})$ assigned to each particle in iteration t. Resampling removes particles with small weights and duplicates particles that have high probability of occurrence. After resampling, the particles are mutated following the MCMC approach, however, the proposed samples are generated from a multivariate Gaussian distribution $K(\bar{\theta}, \Sigma)$

$$p^2.\Sigma_t = p^2 \sum_{i=1}^{N} \frac{w(\theta_{t,i})}{\sum_{i=1}^{N} w(\theta_{t,i})} (\theta_{t,i} - \bar{\theta}_t)(\theta_{t,i} - \bar{\theta}_t)^T \tag{3.5}$$

where $\bar{\theta}_t$ is the weighted mean of the sample at iteration t. It is a vector comprising the means of all the parameters included in the covariance matrix:

$$\bar{\theta}_t = \frac{\sum_{i=1}^{N} w(\theta_{t,i})\, \theta_{t,i}}{\sum_{i=1}^{N} w(\theta_{t,i})} \tag{3.6}$$

The covariance is scaled by the term p^2. The objective is to scale the covariance so that the rejection rate is not too high and at the same time it is convenient well explore the parameter space. Ching and Chen (2007) proposed a value of 0.2 for p. The proposed particle is accepted or rejected based on the acceptance ratio of Metropolis-Hastings as explained previously.

3.2.1.1 CATMIP

In order to better estimate the posterior, several authors updated this general implementation. Minson et al. (2013) proposed in his algorithm CATMIP to run multiple Markov jumps N_{steps} for each chain at each iteration before preceding to the subsequent iteration instead of only doing one jump as in TMCMC. Each chain should be long enough to ensure exploration of the whole parameter space and to ensure that the samples drawn from the priors at the beginning of each chain do not have influence on the posteriors estimated. In TMCMC, proceeding from one iteration to another with only one MCMC jump makes the sampler prone to be trapped in local minimums, which is avoided in CATMIP. Still, if N_{steps} is not sufficient for parameter exploration, CATMIP will also suffer from the same problem. Thus N_{steps} is an essential hyper-parameter of this sampler.

For the same purpose, the scaling factor p is also automatically determined at each iteration to scale the covariance Σ_t as follows:

$$p = w_a + R \cdot w_b \tag{3.7}$$

R is the acceptance rate of the sampler, w_a is the acceptance weight, and w_b is the rejection weight. The idea is if the acceptance rate is high, the covariance is scaled up to explore more the parameter space, otherwise, p scales down the covariance. w_a and w_b are two constants and are set by Minson et al. (2013) to be 1/9 and 1/8 respectively.

Another way to adaptively select β is by ensuring that the effective sample size (ESS) (Eq-3.8) is not less than a predefined threshold which is often chosen to be $N/2$ with N being the total number of samples. ESS measures how informative the sample set is and how independent

or correlated the samples in a chain are. The higher the ESS the more informative and independent the samples are. EES is computed as follows:

$$ESS = \frac{1}{\sum_{i=1}^{N} w(\theta_{t,i})^2} \tag{3.8}$$

The details of CATMIP sampler are illustrated in algorithm 3.2.

Algorithm 3.2

---

1. Sample *N* particles $\theta_t = \{\theta_1, \dots, \theta_N\}$ for $t = 0$ from the prior distributions $p(\theta)$.
2. **while** $\beta_t \leq 1$:
    a. Increment the distribution subscript $t = t + 1$
    b. Estimate β_t such that $cov\{w(\theta_{t,i})\} = 1$ or $ESS = \frac{1}{\sum_{i=1}^{N} w(\theta_{t,i})^2} \geq N/2$
    c. Compute Σ_t and p_t
    d. Resample N particles θ_t^0 from θ_{t-1} with probabilities $w(\theta_{t-1,i})$
    e. **for** i = 1, …, N_{steps}:
        i. Generate a proposal θ_t^i for each chain from $K(\theta_t^i | \theta_t^{i-1}, \Sigma_t^0)$
        ii. Set $\theta_t^{i-1} = \theta_t^i$ with probability $min\left\{1, \frac{f(\theta_t^i|Z)\, K(\theta_t^{i-1}|\theta_t^i, \Sigma_t^0)}{f(\theta_t^{i-1}|Z)\, K(\theta_t^i|\theta_t^{i-1}, \Sigma_t^0)}\right\}$ else retain θ_t^{i-1} unchanged

---

3.2.1.2 SMC variant

Another variant of the SMC family is the algorithm used in Adams et al. (2020). The concept is very similar to CATMIP; however, instead of specifying a number for Markov jumps N_{steps} at each iteration, the sampler keeps on making new jumps until a certain predefined percentage of the samples change their position: 95 % is considered in their paper. In other words, the hyper-parameter N_{steps} in CATMIP is replaced by this percentage. The scaling factor of the kernel covariance is also determined differently. The idea is that at each new Markov jump within a given iteration, the covariance is divided by the square of the jump iterator as follows:

$$K\left(\bar{\theta}, \frac{\Sigma}{r^2}\right) \tag{3.9}$$

where r is the number of MCMC jumps executed in a current iteration. This increases the chance of moving the particles with each new Markov jump. Unlike CATMIP, this approach does not resample at every iteration, however it resamples only when found necessary. That is, the samples are reweighted following equation (3.10), and then normalised. Using these weights, the ESS is computed; if the samples at the previous iteration are found sufficiently independent (ESS largerthan $N/2$), then the samples are moved to the subsequent iteration without being resampled and the weights are updated as follows:

$$w(\theta_{t,i}) = w(\theta_{t-1,i}) \times w(\theta_{t,i}) \tag{3.10}$$

If the samples are found to be correlated, then the particles are resampled with weights and moved to the subsequent iterations and then their weights are reset to one.

The intermediate distributions are specified by selecting β such that:

$$ESS_t = (1 - \Delta)ESS_{t-1} \tag{3.11}$$

where Δ is a number that indicates how close the distributions are to each other. The smaller Δ is the more intermediate distributions are generated and vice versa. Adams et al. (2020) proposed to set: $\Delta = 0.02$. The details of the sampler are illustrated in algorithm 3.3.

Algorithm 3.3

1. Sample *N* particles $\theta_t = \{\theta_1, \dots, \theta_N\}$ for $t = 0$ from the prior distributions $p(\theta)$.
2. Set equal weights $w(\theta_{t,i}) = 1/N$
3. **while** $\beta_t \leq 1$:
    a. Increment the distribution subscript $t = t + 1$ and $r = 1$
    b. Estimate β_t such that $cov\{w(\theta_{t,i})\} = 1$ or $ESS = \frac{1}{\sum_{i=1}^{N} w(\theta_{t,i})^2} \geq N/2$
    c. Reweight all samples $w(\theta_{t,i}) = w(\theta_{t-1,i}) \times w(\theta_{t,i})$ then normalise them
    d. **if** $ESS < N/2$:
        i. Resample $\theta_{t,i}$ from $\{\theta_{t-1,i}, w(\theta_{t-1,i})\}$ and set $w(\theta_{t,i}) = 1/N$
        ii. End
    e. Set number of samples that changed location (N_c) to zero
    f. **while** $100 \cdot N_c/N < 0.95 \cdot N$:
        i. Generate a proposal θ_t^* for each chain from $K(\theta_t^* | \theta_t, \frac{\Sigma_t}{r^2})$
        ii. Set $\theta_t = \theta_t^*$ with probability $min\left\{1, \frac{f(\theta_t^*|Z)\, K(\theta_t|\theta_t^*, \frac{\Sigma_t}{r^2})}{f(\theta_t|Z)\, K(\theta_t^*|\theta_t, \frac{\Sigma_t}{r^2})}\right\}$ else retain θ_t unchanged and denote the number of accepted particles N_{acc}
        iii. Set $N_c = N_c + N_{acc}$
        iv. Increment the MCMC jump iterator $r = r + 1$

3.2.2 Approximate Bayesian computation (ABC)

Approximate Bayesian computation was firstly used in biological sciences by Pritchard et al. (1999). The approach used was lately named rejection ABC. The idea is to approximate the likelihood with a discrepancy function $\rho(.)$ to compute the distance between measurements and model outputs. The framework is illustrated in algorithm 3.4.

Algorithm 3.4

1. Repeat the following until *N* samples are accepted
    i. Draw $\theta_i \sim \pi(\theta)$
    ii. Simulate $y_i \sim p(Z|\theta_i)$
    iii. Reject θ_i if $\rho(S(y_i), S(z)) > \delta$

The introduced tolerance δ is a measure of the accuracy of the algorithm. If the drawn samples yield a difference less than this tolerance, then the sample is considered a sample from the posterior.

The accepted parameter values cannot be considered as samples from the true posterior $p(\theta|Z)$; however, they are samples from another distribution that is an approximate of the posterior $p(\theta|S(Y) - S(Z) \leq \delta)$. The tolerance value determines how close or far this approximation is. A zero tolerance, $\delta = 0$, means that the algorithm is exact and gives draws from the posterior distribution $p(\theta|Z)$ (Beaumont 2010). If δ is very large, the algorithm is inaccurate and all the samples drawn from the prior distribution are accepted so that the posterior PDF is the same as the prior and calibration is useless. A small value of δ leads to a better approximation of the posterior, but it decreases the acceptance rate and thus more computation has to be performed to reach a given sample size of parameter values. A large tolerance value leads to a fast computation but inaccurate results. Consequently, the tolerance δ can be considered as a trade-off between computability and accuracy (Wilkinson 2013).

As stated in the first chapter, there exists different post processing techniques to overcome the considered tolerance. In practice, a higher tolerance value is used for computational efficiency purposes and then, it is accounted for in these post-processing techniques. For more information on this topic, the reader is referred to appendix C.

Marjoram et al. (2003) proposed to use MCMC described earlier within ABC. The algorithm is called ABC-MCMC. In this case, the Metropolis acceptance rejection probability is only applied when the discrepancy between the model output and the observed data is smaller than the identified tolerance. The advantage is that instead of generating all the samples from the prior, a random MCMC walk is performed to detect high probability regions.

ABC-MCMC suffers from the problem of being trapped in low probability regions (Sisson et al. 2007). Normally the sampler will pass onto the tails of the posterior distribution. Thus let us consider the case where a certain parameter value in this tail region is accepted: then the kernel transition function will be centred on this parameter value. Accordingly, the sampler will be obliged to keep moving through the boundaries of this kernel until it finds another accepted value and this will need numerous iterations as it is very unlikely to accept a parameter value in this region. This will cause a high computation inefficiency especially if the starting sample

is chosen far into the tails of the posterior distribution: in MCMC, a random sample is selected to serve as a start for the MCMC chain.

3.2.2.1 ABC-PMC

Sisson et al. (2007) applied the sequential Monte Carlo or "particle filter" concept in the approximate Bayesian computation context to overcome the problem of convergence associated with the previous algorithms and they called it ABC-PRC. Beaumont et al. (2009) introduced another SMC approach named ABC-PMC (population Monte-Carlo). It is based on the PRC approach but with corrected weighting ratios that avoid the bias presented in the PRC method. They showed that the weighting mechanism adapted in PRC is biased and the proposed a correction for it.

The intuition behind ABC-PMC is that instead of drawing one parameter value θ_i at a time, a whole set of N parameter values called particles are drawn from the prior population and assigned weights. Then, each particle is moved in the parameter space to explore it using an appropriate transition kernel and is updated once it is moved to a space that yields an acceptable difference between the measured data and the simulation output based on a chosen tolerance. This can be done T times, where at each iteration, the particles are resampled in a way that the ones that have been assigned higher weights from the previous iteration are kept; and those that are less likely to represent the posterior population are discarded. Thus, in each iteration, particles are resampled from the previous iteration instead of drawing new parameter values from the prior, as it is the case in the basic rejection algorithm.

A second feature of the particle filtering approach is that at each iteration, the tolerance is decreased. In other words, each iteration *t* yields distributions of the input parameters based on the chosen tolerance $P(\theta/S(Z) - S(Y) \leq \delta_t)$, then for the next iteration, this posterior population is updated based on a lower tolerance to better approximate the final posterior population and so on until a certain low tolerance is considered, which closely approximates the true posterior population. This is similar to the SMC applied to the likelihood-dependent algorithm. However, instead of updating β, the tolerance δ is updated and identifies the distribution at each iteration. This enhances the computational efficiency of the approach compared to the previous ABC methods.

In a first step, N parameter values are sampled according to the prior PDF based on the basic rejection ABC algorithm but with a high tolerance value to achieve a high acceptance rate. In this step, equal weights are assigned to all the accepted parameter values. Eventually, in the second iteration, all those chosen particles will be resampled as they are all equally weighted. They are then perturbed based on a certain transitional kernel $K(\theta_i^*|\theta_i^{t-1})$. The sampler keeps on perturbing the particle θ_i^{t-1} until θ_i^* is accepted. Then, a weight is given to this particle. This is done to all the particles resampled from the previous iteration's distribution and thus, a new weighted distribution is obtained. The weighting ratio used in this algorithm is shown in the following equation:

$$w_i^{(t)} = \pi\left(\theta_i^{(t)}\right) / \sum_{j=1}^{N} w_j^{(t-1)} K\left(\theta_i^{(t)} \middle| \theta_j^{(t-1)}; {\sigma_t}^2\right) \tag{3.12}$$

Beaumont et al. (2009) also proposed to evaluate the kernel transition at each iteration based on the previous iteration as follows:

$$K(\theta^*, \sigma_t^2) \tag{3.13}$$

where θ^* is the particle sampled from the previous iteration to be perturbed by the kernel, and σ_t^2 is the variance of the kernel which is proposed by Beaumont et al. (2009) to be twice the variance of the particle of the previous iteration derived this by minimising the Kullback-Leibler divergence between the target and proposal distributions:

$$\sigma_t^2 = 2\sigma_{t-1}^2 \tag{3.14}$$

The kernel function used by Beaumont et al. (2009) is a Gaussian kernel:

$$K(x) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{3.15}$$

where

$$x = \frac{\theta_k^t - \theta_k^{t-1}}{\sigma_t} \tag{3.16}$$

Toni et al. (2009) formulated an equivalent approach to that of Beaumont et al. (2009) but with a different kernel (also called perturbation kernel), which is also automatically adapted

after each iteration. They proposed the use of a normal or uniform distribution with mean θ^* and variance σ_t^2 that depends on the length of a parameter range achieved in population $(t-1)$

In the case of a uniform kernel: $$K_t(\theta|\theta^*) = \theta^* + U(-\sigma_t, \sigma_t) \quad (3.17)$$

In the case of a normal Kernel: $$K_t(\theta|\theta^*) = N(\theta^*, \sigma_t^2) \quad (3.18)$$

where $$\sigma_t = D(max\{\theta\}_{t-1} - min\{\theta\}_{t-1}), \qquad D \epsilon \mathbb{R} \quad (3.19)$$

D is a value that can be changed according to the problem at hand (e.g. 0.5, 1.0, 2.0); the larger *D*, the larger the variance becomes by which more particles fall outside the acceptance region, and this causes a computation inefficiency.

The choice of the perturbation Kernel plays a big role in the computational efficiency of this approach. A local perturbation kernel or a perturbation kernel with a small variance has a high acceptance rate as the candidate (proposed) particle will be close to the previous one. On the other hand, a widely spread out perturbation kernels with big variances explore the parameter space but with a low acceptance rate as the candidate particle will have a high probability of falling outside the acceptance region of the posterior distribution. Thus, the choice of the perturbation kernel and its parameters (e.g. variance, covariance…) is decided as a trade-off between space exploration and the computational efficiency (Filippi et al. 2013). Parameter space exploration is very important to ensure that the samples are drawn from all the range of the posterior in order to well approximate it. If the kernel does not explore sufficiently the parameter space, the samples will only be drawn from a certain region of the posterior and this will lead to a biased estimation of the posterior: the samples approximate a part of the posterior and not all the posterior.

Both kernels used in Beaumont et al. (2009) and Toni et al. (2009) are component-wise perturbation kernels: the particles of each parameter are perturbed independently following an independent kernel specific to each parameter. If the parameter vectors are correlated or have significant interactions, it might get inefficient to perturb the particles based on these kernels, since they do not account for probable correlations and interactions between different parameters. Under these conditions, a multivariate normal perturbation kernels may be more

efficient computationally (Filippi et al. 2013). The multivariate normal function is given as follows:

$$K(x) = \frac{1}{|\Sigma|(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}x} \tag{3.20}$$

where $$x = (\theta_k^t - \theta_k^{t-1})^T \Sigma^{-1} (\theta_k^t - \theta_k^{t-1}) \tag{3.21}$$

As with the previous kernels, the covariance of this kernel at iteration t is based on the previous population parameter covariance. Previously, the kernel is estimated for each parameter separately based on its particles variance in *t-1* ; however, a multivariate kernel is estimated based on the covariance of all the parameter vector in *t-1*. Following the same derivation of Beaumont et al. (2009), Filippi et al. (2013) showed how an optimal covariance for the multivariate kernel can be found to be $\Sigma^t = 2COV\left(\{\theta^{i,t-1}\}_{1\leq i\leq N}\right)$. Consequently in this method, a multivariate kernel, with this optimal covariance is used. Some other kernels are presented in appendix D.

The approach of Beaumont et al. (2009) combines the benefits of the basic rejection and MCMC algorithms where as in the MCMC, the parameter values are drawn from a distribution closer to the posterior instead of being drawn from the prior. As in the rejection algorithm, this approach has no risk to get stuck in a region of low probability. The problem in this approach is in choosing the sequence of tolerance values for each iteration $(\delta_1, \delta_2 \ldots \delta_t)$, and in deciding when to stop the iterations. The sequence of decreasing tolerances can influence the accuracy of the obtained results and the computational efficiency. Moreover, if the tolerance value at the last iteration δ_t is too large, the posterior will not be well estimated, and a very small value could be infeasible.

One solution to this problem is to determine the tolerance of each iteration δ_t based on the samples retained in the previous one. The idea is to complete the first iteration with a predefined tolerance and then for the subsequent iterations, the tolerance δ_{t+1} is chosen by taking the α-quantile of the distances of the previous iteration: δ_{t+1} is the maximum distance below the α-quantile recorded in $\rho_{1\leq i\leq N}^{t-1}$. This is called the quantile approach for deterning the tolerances. The steps of this algorithm are illustrated in algorithm 3.5.

Algorithm 3.5

1. Initialise by setting: a final tolerance δ_T
2. **for** $i = 1 \dots N$:
    a. Sample θ_i from $\pi(\theta)$ and simulate $x \sim f(x|\theta_i)$
    b. Compute and save the distance $\rho_i = \rho\big(S(x), S(y)\big)$
    c. Set the weights $w_i = 1$ and $t = 2$
3. **for** $t = 2 \dots T$:
    a. Let δ_t be the first α quantile of $\rho_i^{t-1} = \left\{\rho_i^{t-1}\right\}_{1 \le i \le N}$
    b. Compute the variance σ_{t-1}^2 of $\left\{\theta_i^{(t-1)}, w_i^{(t-1)}\right\}$ and set $\sigma_t^2 = 2\sigma_{t-1}^2$
    c. **for** $i = 1 \dots N$:
        i. Sample θ^* from $\left\{\theta^{(t-1)}\right\}$ with weights $\left\{w^{(t-1)}\right\}$
        ii. Generate $\theta' \big| \theta_i^* \sim K\big(\theta_i^*, \sigma_{(t-1)}^2\big)$ and simulate $x \sim f(x|\theta')$
        iii. **if** $\rho\big(S(x), S(y)\big) \le \delta_t$:
            - Set $\theta_i = \theta'$ and increment $i = i + 1$
    d. Set the weights $w^{(i,t)}$ and normalise them and increment $t = t + 1$

3.2.2.2 APMC

There are different methods to implement the quantile approach. All such algorithms follow the same principle of determining δ_{t+1} as the maximum distance below the α-quantile recorded in $\rho_{1 \le i \le N}^{t-1}$; however with some differences in their implementation. Lenormand et al. (2013) implemented a different method to apply the quantile approach. They modified the framework in ABC-PMC algorithm and they called their algorithm adaptive population Monte-Carlo ABC (APMC). They used the same perturbation kernel proposed by Beaumont et al. (2009) with σ_t^2 taken as twice the variance of the previous iteration samples. The weight expression for the particles at iteration t differs from that of the ABC-PMC algorithm by the addition of $\sum_{k=1}^{N_\alpha} w_k^{t-1}$ in the denominator as shown in the following expression:

$$w_i^t = \frac{P(\theta_i^t)}{\sum_{j=1}^{N_\alpha} \left(\frac{w_j^{t-1}}{\sum_{k=1}^{N_\alpha} w_k^{t-1}} \right) K\left(\theta_i^t / \theta_j^{t-1}; \sigma_t^2\right)} \tag{3.22}$$

This formula is used to weight the new particles $\theta_{N_\alpha \dots N}$ that are generated on top of $\theta_{1 \dots N_\alpha}$. The reason behind this variation is that in this algorithm, instead of generating the sample at a

new iteration from scratch, new particles are generated and concatenate with the particles of the previous sample set. Thus, the scaling of weights needs to be consistent across the different steps of the algorithm (Lenormand et al. 2013).

At the first iteration, N particles are drawn from the prior distribution and the model simulates the corresponding $x \sim f(x|\theta_i)$. The particles are weighted and the discrepancy distance $\rho(\cdot)$ for each particle is saved as shown in algorithm 3.5. The tolerance δ is then taken to be the first alpha-quantile of these distances. The particles chosen for this iteration are those that yielded a tolerance lower than the calculated one, and their number is noted as N_α.

In the next iterations, a candidate particle θ^* from the previous sample is chosen according to its weight w_i^t and then perturbed with a transitional kernel. The discrepancy of the newly generated particle is saved and a weight is assigned to the new based on equation (3.22). This is done to $N - N_\alpha$ particles. The new simulated particles after perturbation are concatenated with the ones from the previous iteration N_α with their weights and distances to form N particles. The tolerance δ of this new iteration can now be determined based on α-quantile of the discrepancy distances of all the previous and the concatenated particles. The N particles, ensemble of previous iteration N_α and the ones attained after perturbation in this iteration $N - N_\alpha$, are filtered according to the newly indicated tolerance δ. The chosen particles form the sample of this current iteration t and their number is noted N_α again.

The difference is that in the regular ABC-PMC, each particle in the previous iteration is perturbed until it is moved to a value that yields a lower discrepancy than a predefined tolerance and a new sample is formed that only contains the new particles. However, in this algorithm, each particle is perturbed only once and then weighted to be subjected to filtering after the indication of the appropriate tolerance.

Lenormand et al. (2013) also introduced a stopping criterion. At each iteration, the proportion of the accepted particles among the $N - N_\alpha$ new particles are calculated:

$$p_{acc}(t) = \frac{1}{N - N_\alpha} \sum_{k=N_\alpha+1}^{N} \mathbb{I}_{\rho_k^{(t-1)} < \delta_{t-1}} \tag{3.23}$$

The tolerance and the discrepancy are subscripted with $t - 1$ to indicate that the particles are concatenated with the ones from the previous iteration. If this proportion is below a

predetermined threshold$p_{acc_{min}}$, the algorithm terminates. This chosen criterion ensures that additional simulations would not have considerable changes on the posterior distribution. They proved that the algorithm will terminate even if $p_{acc_{min}} = 0$, this ensures that the algorithm converges. The details of this method are shown in algorithm 3.6.

Algorithm 3.6

---

1. Initialise the algorithm with $N, N_\alpha = \alpha N, \alpha \in \lfloor 0,1 \rfloor$, and $p_{acc_{min}}$.
2. **for** $i = 1, \dots, N$:
    a. Generate a sample from the prior $\theta_i^{(0)} \sim \pi(\theta)$
    b. Save the distances $\rho_i^{(0)} = \rho\big(S(x), S(y)\big)$
    c. Set the weights $w_i^{(0)} = 1$
3. Let δ_1 be the first α quantile of $\rho^{(0)} = \left\{\rho_i^{(0)}\right\}_{1 \le i \le N}$
4. Let $\left\{\theta_i^{(1)}, w_i^{(1)}, \rho_i^{(1)}\right\} = \left\{\left(\theta_t^{(0)}, w_t^{(0)}, \rho_t^{(0)}\right) \middle| \rho_i^{(0)} \le \delta_1, 1 \le i \le N\right\}$
5. Compute the variance σ_1^2 of $\left\{\theta_i^{(1)}, w_i^{(1)}\right\}$ and set $\sigma_1^2 = 2\sigma_1^2$
6. Increment $t = 2$ and $p_{acc} = 1$
7. **while** $p_{acc} > p_{acc_{min}}$:
    a. **for** $i = N_\alpha, \dots, N$:
        i. Sample θ_i^* form $\theta_j^{(t-1)}$ with probability $\frac{w_j^{(t-1)}}{\sum_{K=1}^{N_\alpha} w_k^{(t-1)}}, 1 \le j \le N_\alpha$
        ii. Generate $\theta_i^{(t-1)} \big| \theta_i^* \sim K\big(\theta_i^*, \sigma_{(t-1)}^2\big)$ and simulate $x \sim f\left(x \middle| \theta_i^{(t-1)}\right)$
        iii. Set $\rho_i^{(t-1)} = \rho\big(S(x), S(y)\big)$
        iv. Set the weights $w_i^{(t-1)} = \frac{\pi(\theta_i^*)}{\sum_{J=1}^{N_\alpha}\left(w_j^{(t-1)} / \sum_{k=1}^{N_\alpha} w_k^{(t-1)}\right) K\left(\theta_i^* / \theta_j^{t-1}; \sigma_{(t-1)}^2\right)}$
    b. Compute p_{acc} from equation (3.23)
    c. Let δ_t be the first α quantile of $\rho^{(t-1)} = \left\{\rho_i^{(t-1)}\right\}_{1 \le i \le N}$
    d. Let $\left\{\left(\theta_i^{(t)}, w_i^{(t)}, \rho_i^{(t)}\right)\right\} = \left\{\left(\theta_i^{(t-1)}, w_i^{(t-1)}, \rho_i^{(t-1)}\right) \middle| \rho_i^{(t-1)} \le \delta_t, 1 \le i \le N\right\}$
    e. Compute the variance σ_t^2 of $\left\{\theta_i^{(t)}, w_i^{(t)}\right\}$ and set $\sigma_t^2 = 2\sigma_t^2$
    f. Increment $t = t + 1$

3.2.3 Machine learning for calibration

A new class of Bayesian inference that belongs to the approximate Bayesian computation is proposed by Raynal et al. (2017). It is based on using the machine learning framework in the context of ABC. They proposed to use the random forest RF for parameter inference. A brief introduction on random forests is firstly presented in this section and then its relation to Bayesian calibration is clarified.

3.2.3.1 Random forest concept

Random forest is a supervised machine learning algorithm used for classification and regression. It is an ensemble learning algorithm where it combines the outputs or the classification decisions of individual classifiers called decision trees (Yan and Goebel 2004). Decision tree is a predictive model that relates the input (predictors θ) of a certain process to its outputs (responses y) by following a series of splitting decision to the predictor space. It starts with the full predictor data set available which represents the tree base and splits it into two parts called nodes based on a splitting rule. The tree base is called a root node. Each node m is then split following the same methodology until no more than N_{min} samples are categorised in the node. Another stopping criterion is when all the samples in a given node have the same response value or the same predictor value. The final nodes from which no further nodes are generated are called leaf nodes. For regression N_{min} is often set to 5, and the average $\bar{y}_m$ of the samples n_m responses in the leaf node represents the prediction value associated with this node.

$$\bar{y}_m = \frac{1}{n_m}\sum_{i}^{n_m} y_i \tag{3.24}$$

For continuous problem as it is the case for BEMs, the splitting rule is a comparison held on a parameter selected in the predictor space:

$$\theta_j \leq v \text{ or } \theta_j > v \tag{3.25}$$

where j is a subscript that indicates the parameter chosen for the splitting decision among the predictor space. The best parameter j and the best value of v for a split decision at a given node m are selected by minimising the following equation:

$$\frac{n_{m_r}}{n_m} I(m_r) + \frac{n_{m_l}}{n_m} I(m_l) \tag{3.26}$$

where n_{m_r} and n_{m_l} are the number of samples present in the right and left child nodes respectively. $I(.)$ is the criterion applied to each child node: the two nodes split from one node. For classification, there are different criteria that can be applied such as the Gini impurity or entropy. For regression, the L_2 loss function (sum of squared errors) is often applied:

$$I(m) = \frac{1}{n_m}\sum_{i=1}^{n_m}(y_i - \bar{y}_m)^2 \tag{3.27}$$

where $\bar{y}_m$ is the average of the responses of all the samples in the node. This weights the loss function of each child node based on the number of samples in each one relative to node m. The combination of (j, b) that minimises equation (3.26) is selected.

After the decision tree is trained on the available data set, the prediction of a new sample θ^* is achieved by applying the splitting decisions of the trained tree on the sample until it reaches a leaf node. The prediction of the sample $f(\theta^*)$ is then the value associated with that leaf node. A general name that refers to both classification and regression is classification and regression tree (CART) after Breiman et al. (1983).

Breiman (2001) based on previous developments in the field: (Amit and Geman 1997; Breiman 1996; Ho 1998; 1995) and some novel ideas introduced a method for building a random forest. Random forest is an ensemble classification algorithm that combines the predictions of B uncorrelated trees by taking their average for regression as shown in equation (3.28) where b is the subscript indicating the individual decision tree in the ensemble. For classification, the class corresponding to the maximum votes among the B trees is the final prediction of this ensemble predictions.

$$\mu = \frac{1}{B}\sum_{b=1}^{B}\mu_b(\theta^*) \tag{3.28}$$

With random forests, each tree is trained on a bootstrap sample of the original data set with the same sample size N. Bootstrapping dates back to Efron (1979). It is a resampling technique where samples are randomly generated with replacement from a set of data.

Another aspect of random forest is that the minimisation of equation (3.26) is performed on a subset of parameters n_{try} uniformly drawn from the whole predictor space. Each time a node needs to be split a new sample n_{try} is drawn. Different studies have been performed to analyse the effect of the random forest hyper-parameters (Biau and Scornet 2015; Genuer et al. 2010; Genuer et al. 2008). Here, the values for the hyper-parameters recommended by Raynal et al. (2017) that comply with the general recommendations in RF context are retained. The general steps followed to construct the random forest are shown in algorithm 3.7.

Algorithm 3.7

1. Generate the data set $N = \left(\theta_{1:P,1}, \dots, \theta_{1:P,N};\ y_1, \dots, y_N\right)$ comprising perdictors and responses.
2. **for** b in $\{1, \dots, B\}$:
   a. Generate a bootstrap sample N_b from N
   b. **for** $n_m > N_{min}$ and other stopping criteria not met:
      i. Split node m into m_r and m_r
      ii. Draw parameter subset $\theta_{ntry} = \left(\theta_{1,\dots,n_{try}}\right)$ from predictor space $\theta = (\theta_1, \dots, \theta_P)$
      iii. **for** $i = 1, \dots, n_{try}$:
         - Find bound b_i that minimises equation (3.26)
      iv. Find the combination (θ_i, b_i) among θ_{ntry} that minimises equation (3.26)
      v. Set the combinations as the splitting rule for node m
   c. Associate the average value of the responses in the leaves as the predictions of the leaves

3.2.3.2 Random forest ABC (ABC-RF)

Pudlo et al. (2015) firstly proposed to use the random forest classifier to replace ABC for model selection. They called it ABC-RF. Raynal et al. (2017) then extended the application of ABC-RF so that it is not only used for model selection but also for parameter inference. In this case, the random forest is trained for regression instead of classification. Normally RF is used as a machine learning algorithm to generate predictions *y*: which can be the heat consumption of a building, temperature profile or its summary statistics. Accordingly, the building parameters θ are the predictor space on which the splitting rules and decisions will be performed to construct the random forest, and y are the predictions associated with each corresponding leaf. With ABC-RF , the aim is to estimate the parameters θ given the data *y*. Thus, the predictions of the random forest associated with each leaf will be in this case the predictions of the parameters θ, and the predictor space on which the splitting rules will be applied to contruct the random forest will be *y* or its summary statistics..

ABC-RF has several advantages against ABC algorithms in that it overcomes the difficulty in choosing appropriate summary statistics for the data available which is one of the main problems in ABC algorithms. This is related to the characteristic of random forest: it can handle numerous predictors (summary statistics of responses y in this case) including some that are totally uninformative. This is not very critical with BEM applications in the case of time

series data, where the RMSE could be a sufficient summary statistic. Another main advantage is that it does not require to define a distance function or a minimum threshold. The classification and splitting rules taken sequentially throughout the construction of each tree approximate well the likelihood function and thus do not need a distance metric. It is worth mentioning that each random forest solves the inference problem for one parameter and it is required to train separate random forests for each parameter. In this framework, the objective is to extract statistics that describe the parameter posterior: posterior mean, posterior variance, posterior quantiles.

To apply ABC-RF, a data set of N samples comprising the parameters to calibrate $\theta = (\theta_{1:P}^1, \dots, \theta_{1:P}^N)$ and their corresponding outputs $y = (y^1, \dots, y^N)$ is generated. The outputs y can be replaced with summary statistics describing them. This ensures that the interactions between the parameters are taken into account. Then individual RFs are trained to infer each parameter separately. This means thatthe data set on which the random forest $RF(\theta_i)$ is trained to estimate the parameter θ_i comprise only the samples of this parameter with their corresponding BEM outputs $\{\theta_i^1, \dots, \theta_i^N; y^1, \dots, y^N\}$. Another thing to point out is that in the case of using RF for parameter inference, the parameter space will be called the response variable and the outputs are the predictors. Lets denote by Y the summary statistics of the data Z, and by $l_b(Y)$ the leaf in tree b in which the data Y is categorised after following the splitting rules of $RF(\theta_i)$. The parameter value of θ_i corresponding to data Y predicted by tree b is computed as follows:

$$\mu_b(Y) = \frac{1}{n_{l_b}} \sum_{j=1}^{N} (n_b^j 1_{\{y_i^j \in l_b(Y)\}})\, \theta_i^j \tag{3.29}$$

This expression corresponds to the weighted average of the responses $\theta_i^{1:N}$ belonging to leaf $l_b(Y)$ where Y is categorised. n_{l_b} is the number of samples in leaf $l_b(Y)$ given by:

$$n_{sl_b} = \sum_{j=1}^{N} (n_b^j 1_{\{y_i^j \in l_b(Y)\}}) \tag{3.30}$$

where n_b^j defines how many times a sample $(\theta_i^j; y^i)$ is repeated in the bootstrap used for tree b. This ensures to count all the samples in the leaf including the duplicates. Equation (3.29) can be reorganised as follows:

$$\mu_b(Y) = \sum_{j=1}^{N} w_b^j(Y)\ \theta_i^j \tag{3.31}$$

$$w_b^j(Y) = \frac{n_b^j 1_{\left\{y_i^j \in l_b(Y)\right\}}}{n_{l_b}} \tag{3.32}$$

The responses allocated for all the decision trees of random forest B are then averaged to approximate the posterior expected value of parameter θ_i. The weights are averaged over all the trees for a given sample $\left(\theta_i^j; y^i\right)$:

$$w^j(Y) = \frac{1}{B}\sum_{b=1}^{B} w_b^j(Y) \tag{3.33}$$

The posterior expected values are then given as follows:

$$\mu(Y) = \sum_{j=1}^{N} w^j(Y)\ \theta_i^j \tag{3.34}$$

The steps to estimate the posterior expected value can be combined as follows:

$$\mu(Y) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{N} \frac{n_b^j 1_{\left\{y_i^j \in l_b(Y)\right\}}}{n_{l_b}}\ \theta_i^j \tag{3.35}$$

Equation (3.35) averages the response of each leaf; then, the average over all the trees B is retained. The weighted variance of the posterior can be approximated as follows:

$$V(Y) = \sum_{j=1}^{N} w^j(Y)\left(\theta_i^j - \mu(Y)\right)^2 \tag{3.36}$$

Raynal et al. (2017) proposed another variance estimator that is more specific to the posterior distribution. They proposed to use the output of bag (OOB) samples for the variance estimation: the samples that are not included in the bootstrap of a given decision tree. Those OOB samples are then allowed to follow the splitting rules of the trees in which they are not

used in the process of training. The predictions are then averaged over all these trees and are represented as $\mu_{OOB}(y)$. The estimator is given as follows:

$$V(Y) = \sum_{j=1}^{N} w^j(Y) \left(\theta_i^j - \mu_{OOB}^j\left(y^j\right)\right)^2 \tag{3.37}$$

3.3 Methodology and criteria

The methods presented in the previous section (ABC-PMC, APMC, Adams, CATMIP, ABC-RF) are compared in terms of accuracy and computational efficiency. The comparison is performed following two criteria. The first is the method ability in converging to the parameters true values. The second is the methods' ability to fit accurately to the generated virtual data.

The first criterion is evaluated by computing the normalised Euclidean distance between the posteriors samples and the true values of the parameters as follows:

$$d_{dist} = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{N_s}\left(\frac{\theta_i^j - \theta_{true}^j}{\theta_{true}^j}\right)^2} \tag{3.38}$$

where N_s is the number of samples drawn from the posterior, k is the number of parameters, and θ_{true} is the true value of the parameter used for normalisation. This distance function is applied to each parameter. The smaller the distance the closer the posterior is to the true value and vice versa.

The parameters on which this criterion is applied do not include a temperature parameter, otherwise, the normalisation would not be valid. If a temperature parameter is to be calibrated, then another criterion would be more convenient such as the one used by Juricic (2020). This criterion assumes a ± 5 % error around the true values which is shown in grey in Figure 3.1. Then the integral of the density function lying in this ± 5 % acceptability error is computed and represents how close or far the distribution is from the true value. This criterion is depicted in Figure 3.1.

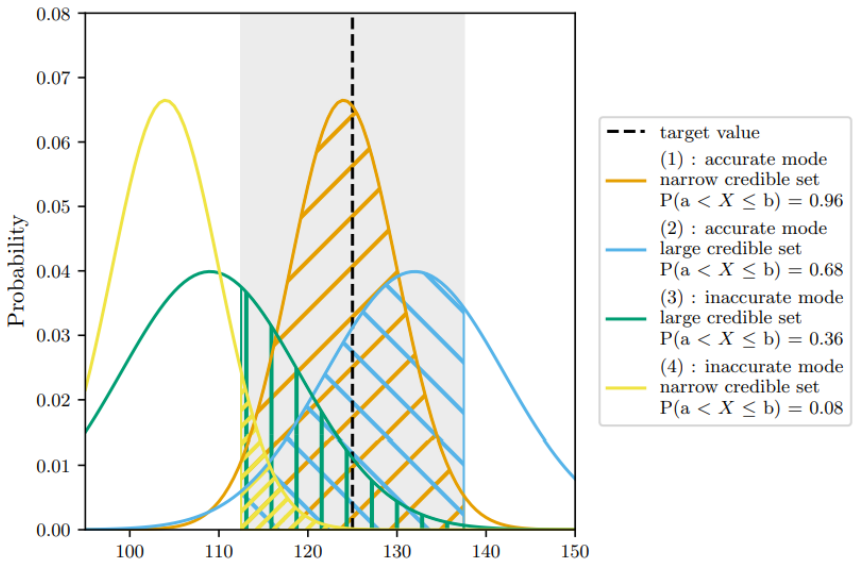

Figure 3.1: Graphical explanation of the citerion taken from Juricic (2020)

The second criterion is to validate the performance and the prediction accuracy of the calibrated model, the priors and the posteriors of the parameters are propagated. This provides information regarding the uncertainties in the prediction before and after calibration. To quantify the model predictive performance, the temperature root mean square error (RMSE) between the measurements and the predictions is evaluated for each propagated sample and the average of all the samples is retained. Other indicators such as the mean absolute error could have also been used.

These criteria are not only evaluated on the posterior distributions, but also on the distributions generated at each iteration of the algorithms. This enables to evaluate the performance of each algorithm with an increasing number of model evaluations which allows for a more comprehensive comparison.

3.4 Results

The calibration methods in this section are applied to the case study presented in chapter 2. Accordingly, following the results of the sensitivity analysis in the previous chapter, the parameters selected for calibration are: ventilation flowrate ($\dot{V}$), heating power (Q_p), specific heat of wall concrete $(c_{p,concW})$, conductivity of polystyrene wallmate (λ_{polW}), dissipated heat (Q_d), and solar albedo (Alb). These are the same parameters calibrated by Robillart (2015) and accordingly, the same priors are retained here as listed in Table 3.1.

Table 3.1: Prior distributions (Robillart 2015)

Parameters	Distribution	Mean	σ	Unit
Ventilation flowrate	Normal	110	11	$[m^3/h]$
Dissipated heat	Normal	208	20.8	$[W]$
Heating power	Normal	1200	20	$[W]$
Concrete specific heat	Normal	2120	212	$[J/(m^3.K)]$
Solar albedo	Normal	0.35	0.035	$[-]$
Conductivity polys	Normal	0.03	0.003	$[W/(m.K)]$

Figure 3.2 shows the model predictive performance of each algorithm averaged over all the scenarios against the number of model evaluations with a simulation budget of 30,000. APMC and Adams showed a similar performance; they converged better and faster to lower RMSE values than CATMIP and ABC-PMC, the latter being the slowest to converge. ABC-RF shows a different performance. Contrarily to the other algorithms, its accuracy with increasing model evaluations did not show a continuous increase: the RMSE indicator slightly decreased with higher model evaluations with significant variability. It is important to mention that recommended default values of the algorithms' hyper-parameters are selected. For instance the size of leaves in ABC-RF is taken as recommended by Raynal et al. (2019); it was not changed with increasing number of simulations.

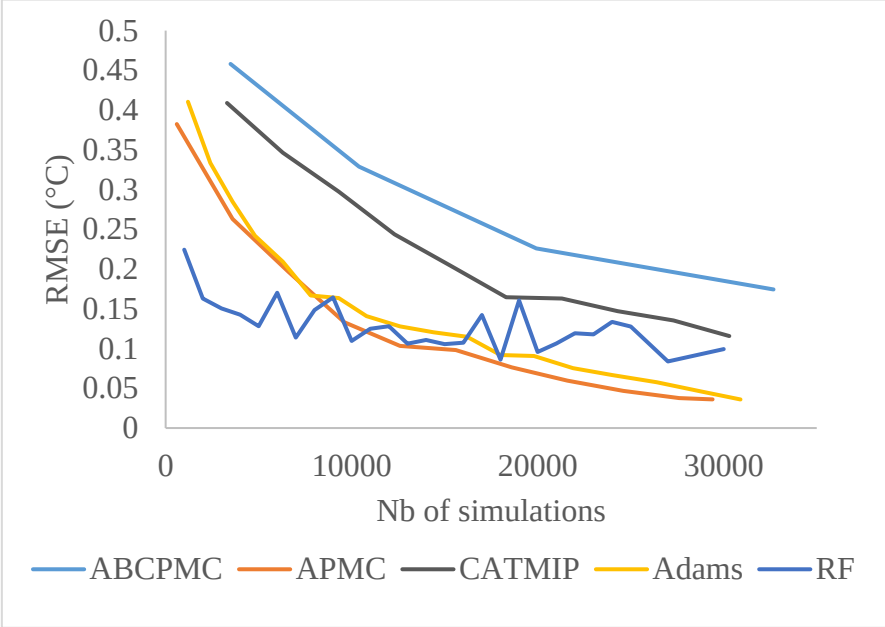

Figure 3.2: Model prediction accuracy for each algorithm (virtual data)

Since the true values of the parameters are known, it is also easy to evaluate the estimation of the parameters by computing the distance between the distribution and the posteriors. Accordingly the Euclidean distance d is computed for each algorithm at each iteration as described in section 3.3. Figure 3.3 shows the Euclidean distance for each algorithm for a simulation budget of 30000. It is clearly depicted that APMC and Adams estimate the true values of the parameters more accurately than the other methods with the specified simulation budget. ABC-PMC shows the least accurate estimation of the true values. ABC-RF estimates the parameters better with less model evaluations, however, it does not show a better estimation with increasing number of model evaluations.

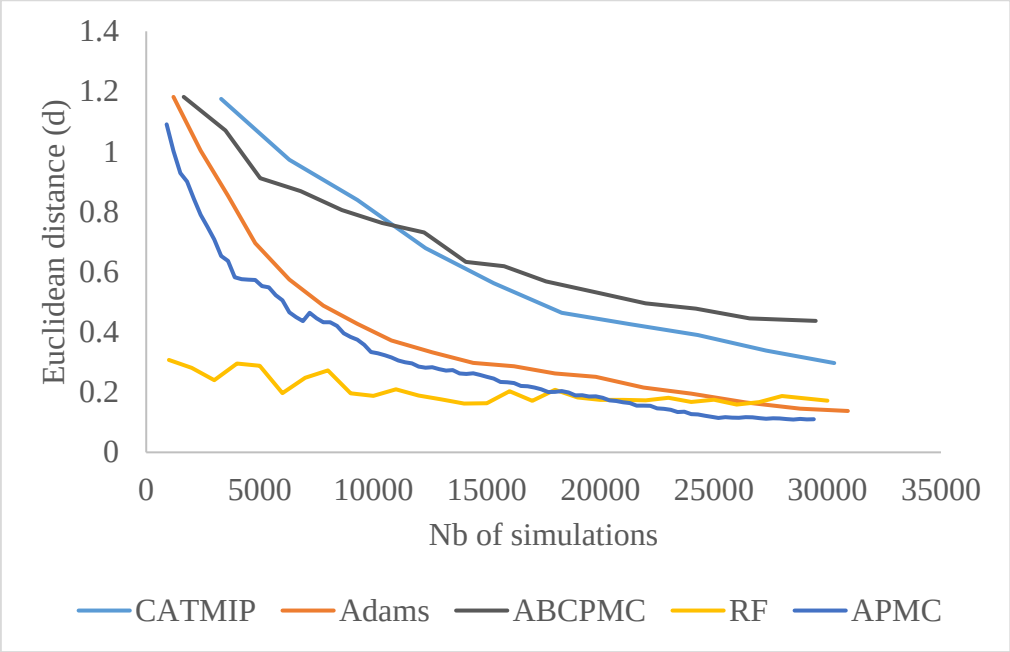

Figure 3.3: Euclidean distance at each iteration

Another difference is that all the algorithms except ABC-RF are capable of reaching similar accuracies in finding the true values of the parameters and in model prediction performance but with different numbers of model evaluations. CATMIP and ABC-PMC required 54,000 and 95,000 simulations respectively (not shown here) to reach the same accuracies (same RMSE of 0.035 °C) attained by Adams and APMC with only 30,000 simulations. However, the model predictive accuracy of ABC-RF is less than the others even when trained with a data set size of 100,000 samples: RF posterior yielded an RMSE of 0.08°C.

These differences could be related to the difference in the criteria of convergence of each iteration between the algorithms. ABC-PMC discards all the samples of a current iteration when

switching to the subsequent one. That is, at each iteration the algorithm keeps on sampling until N samples are accepted for the current iteration threshold. In this case study, it required between $2 \times N$ and $8 \times N$ samples in each iteration all of which were discarded in the next iteration. On the one hand, this is beneficial for a better parameter space exploration, on the other hand it makes the algorithm computationally intensive. APMC avoids this problem by keeping the particles accepted in the current iteration and dragging them to the next iteration. In the next iteration, only $N/2$ particles are sampled randomly with weights then perturbed and evaluated by the model no matter whether they yielded the corresponding threshold or not. Consequently, a larger number of iterations are required to reach convergence but with less simulations per iteration. ABC-PMC required 35 iterations to converge, whereas APMC required 96 iterations but with less simulations in total.

Adams outperformed CATMIP in terms of computational efficiency for the current example with the specified hyper-parameters. The main difference between both algorithms is that Adams automatically specifies the number of required simulations per iteration. It keeps on sampling new particles until a pre-specified percentage of the whole pool has moved to a better location given the current iteration properties and distribution. On the contrary, CATMIP walks a chain at each iteration with a predetermined number of jumps N_{steps}. If the number of steps, N_{steps}, is not wisely defined, the algorithm might either collapse due to particles degeneracy or generate too many unneeded samples. Particles degeneracy usually occurs in sequential Monte Carlo sampling techniques. It describes the problem when all the particles in an iteration collapse to a single particle. In SMC samplers applied to Bayesian inference, it mainly happens when there are not enough MCMC jumps. From this analysis, it could be confirmed that Adams is more robust in terms of its hyper-parameters than CATMIP, but a final conclusion around which one outperforms the other cannot be drawn.

Figure 3.4 shows the cumulative number of model evaluations during the evolution of the algorithms. It clarifies the differences in the simulations required to move from one iteration to another for all the algorithms.

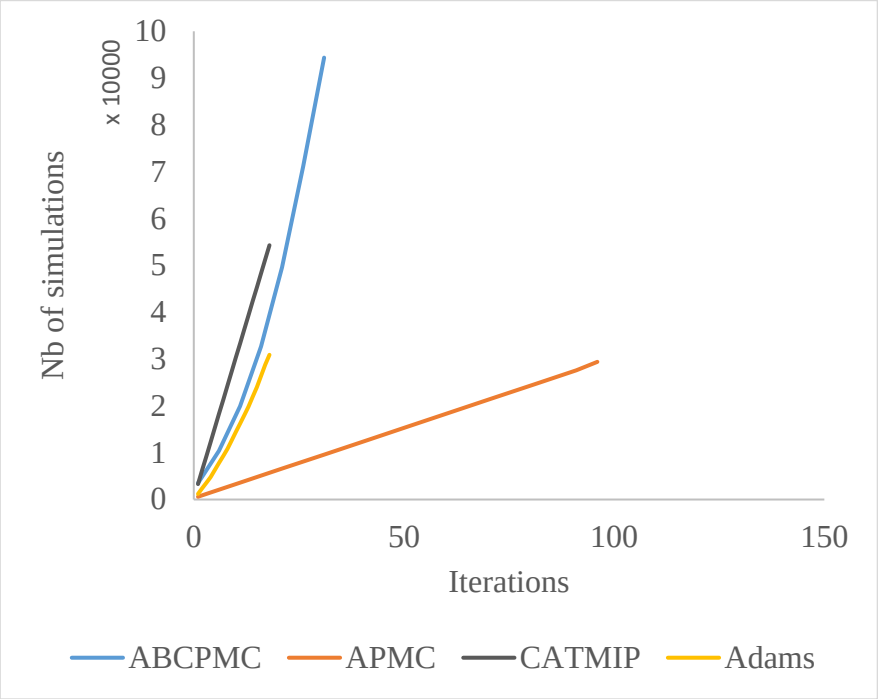

Figure 3.4: Model evaluations with evolution of iterations

Moreover, ABC-RF finds the area around the true value even with only 1000 simulations. Figure 3.3 shows only a comparison between the methods about their precision in finding the true values, and it does not provide a clear idea about how close the estimated distributions are to the true values. Figure 3.5 shows the parameters estimated by ABC-RF and APMC with only 1000 model evaluations. APMC is selected for this comparison since it is shown in Figure 3.3 that it is the faster in estimating the true values compared to the rest. Figure 3.5 shows that ABC-RF is capable with this relatively small data set to move the priors towards the true values regions. For some parameters, the estimated distributions are centred very close to the true values. On the contrary, this is not observed with APMC, where the estimated distributions are closer to the priors than to the true values, which means that it needs more simulations to be able to move the priors closer to the posteriors. This means that ABC-RF is capable of exploring the parameter space better and faster than the other algorithms.

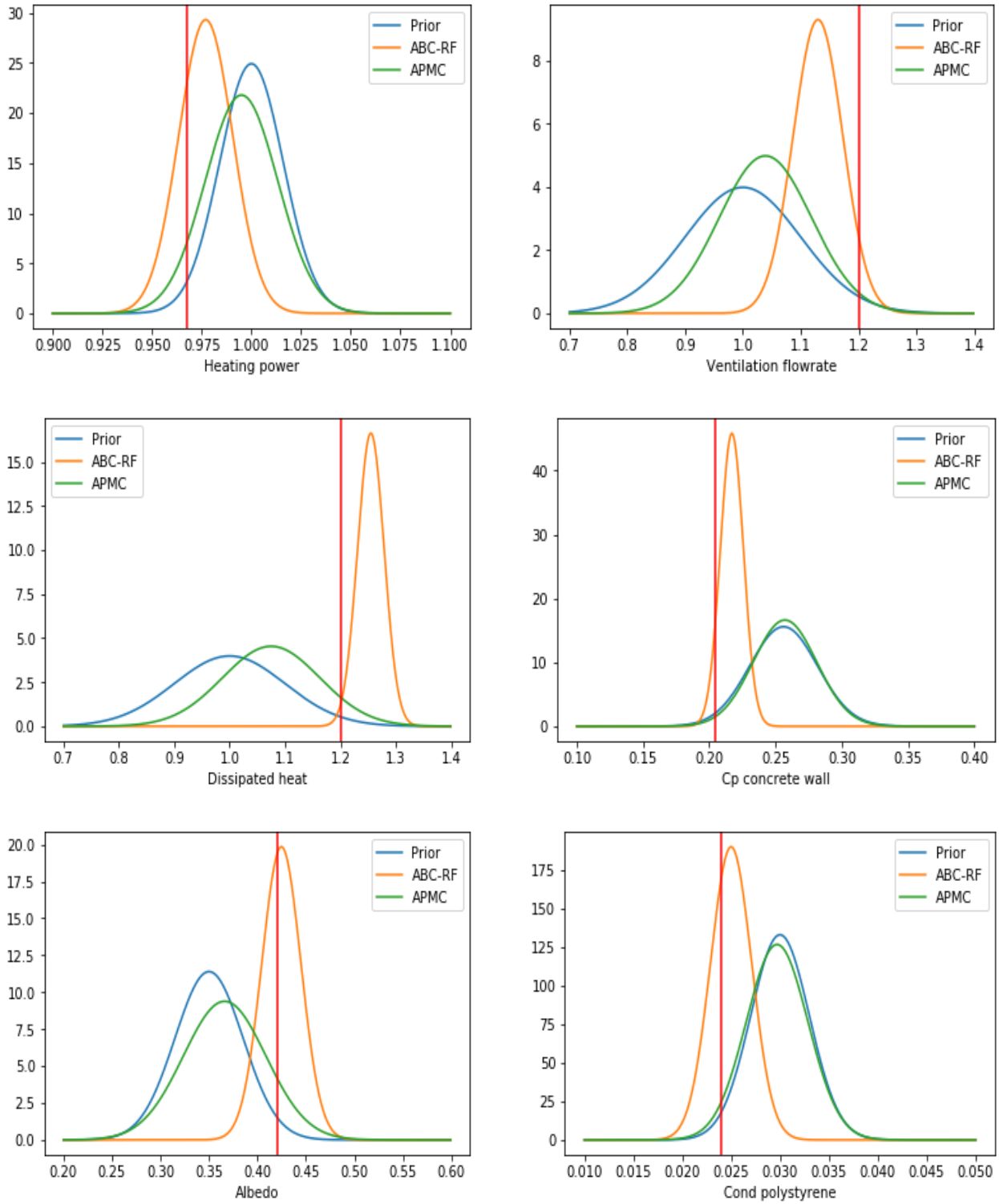

Figure 3.5: Parameters estimation after 1000 model evaluations

Unlike the other algorithms, it is observed that ABC-RF is not able to narrow the posteriors towards the true values with a larger data set. With 30000 as a simulation budget, Figure 3.6 shows the posteriors attained by ABC-RF and APMC. It is clearly depicted that even if the

posteriors of ABC-RF are close to the true values, they are still wider than those of the posteriors obtained by APMC.

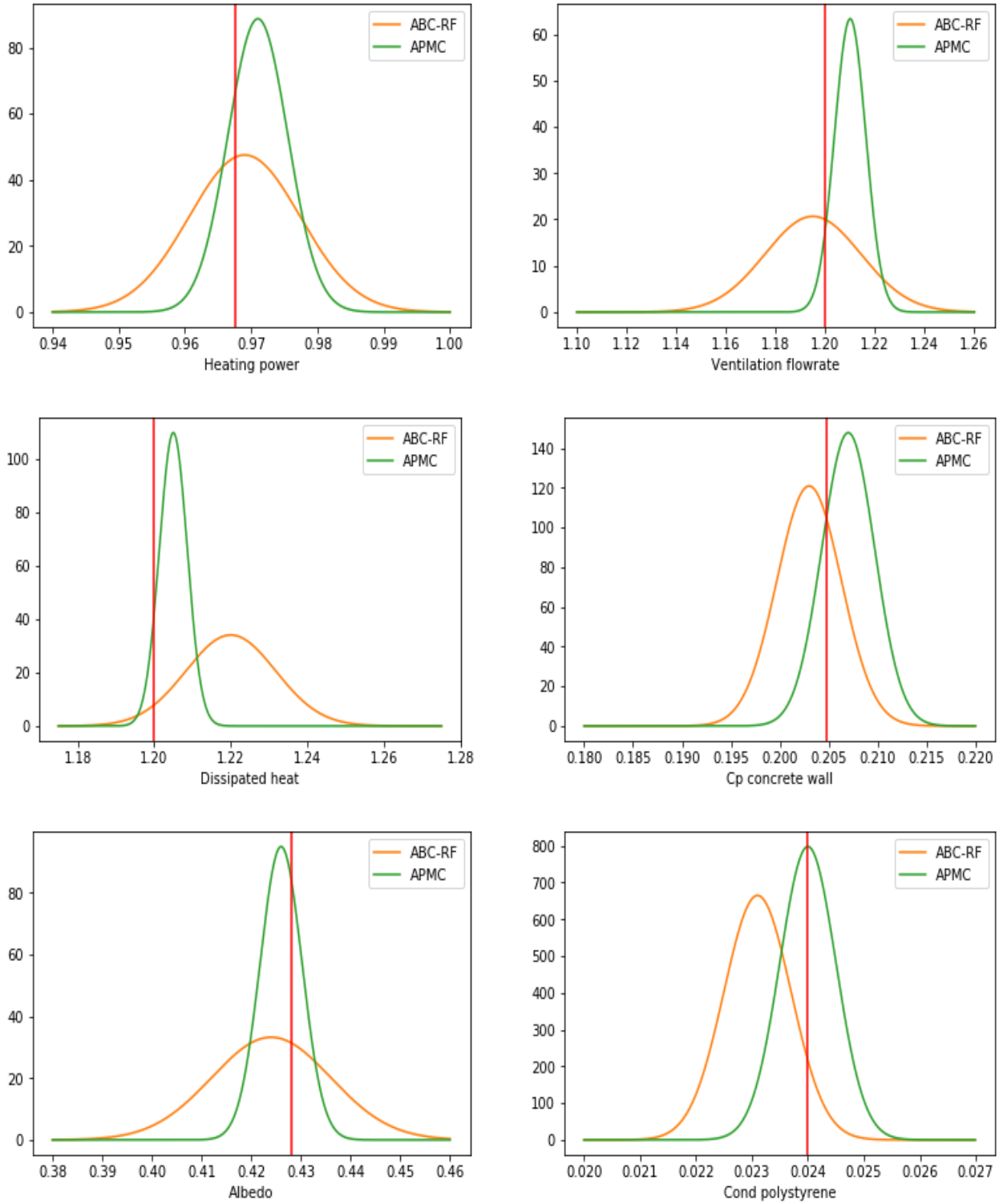

Figure 3.6: Parameters estimation after 30000 model evaluations

The random forest in ABC-RF is trained on all the samples generated from the priors, the samples close and far from the true values. On the contrary, within the other algorithms, the

samples are generated sequentially from distributions closer to the posteriors than to the priors, and at each iteration, the samples in the low probability regions are discarded in favour of those in the high probability region. This could explain the reason behind the wider posteriors obtained by ABC-RF.

Random forests are known for their randomness and this is what has been confirmed in this application where it shows significant variabilities as in Figure 3.2. However, from the same figure, it is depicted that even with the variabilities weighing on the method, it performs better than the other methods with a small number of model evaluations: less than 10000. It means that there is a potential in RF with small data set size, however, the rest can attain better accuracies with increasing model evaluations.

3.5 Conclusion

Calibration of building energy models has recently attracted the focus of researchers in the field especially the application of Bayesian approaches. A significant work has been dedicated to these approaches in order to enhance their performance in terms of accuracy, robustness, and computational efficiency. In this chapter, different Bayesian methods are applied on a virtual data to assess their performance in terms of precision and computational efficiency. Five algorithms were selected, according to their popularity and their ability to be parallelised.

The analysis revealed that even if the likelihood function is approximated by a distance metric in the approximate Bayesian computation algorithms, it does not mean that they perform worse than the likelihood-dependent approaches. It is shown that all tested algorithms yielded sufficient accuracy for the training and the testing data no matter what group they belong to. APMC and Adams outperformed the other algorithms in terms of computational efficiency and precision. The different samplers integrated in each algorithm are the key differences. ABC-RF gave the least precise results with larger data sets and the most precise ones with small data sets. It would be interesting to increase the leaves size and number of trees with increasing number of simulations. This might yield a better convergence behaviour for ABC-RF with more simulations.

Due to its relatively better performance with limited number of model evaluations, ABC-RF might be a better choice if approximate results are needed. If more precise fit to the data is

needed, then APMC or Adams methods could be a better choice. These results orient the efforts towards more investigation on ABC-RF which is presented in chapter 4.

Chapter 4

Identifiability analysis

Identifiability between the parameters is essential when calibrating a model. Due to un-identifiability, the model may not converge to unique solutions. Therefore, it is important to evaluate the identifiability between the model parameters before calibration. This chapter presents an identifiability method and applies it on a virtual case study. The objective is to assess its effect on the calibration methodology. In the second part of the chapter, the effect of increasing number of parameters on the calibration result is assessed.

Résumé du chapitre

L'importance de calibrer un sous-ensemble des paramètres du modèle énergétique du bâtiment a été clarifiée dans les chapitres précédents. Dans le chapitre 2, les méthodes de calibrage sont évaluées suite à une analyse de sensibilité qui classe les paramètres par ordre d'importance. L'ensemble des paramètres les plus importants qui sont généralement pris en compte pour le calibrage peut ne pas correspondre aux paramètres les plus estimables : une interaction significative peut exister au sein des paramètres les plus influents, ce qui rend l'identification plus difficile. De plus, des problèmes d'identifiabilité peuvent survenir en raison de données non informatives ou insuffisantes. L'analyse d'identifiabilité permet de sélectionner les paramètres qui sont les plus influents et qui ont le moins d'interaction. L'effet des données sur l'identifiabilité n'est pas considéré dans ce travail.

Dans ce chapitre, basé sur la littérature, une méthode d'identifiabilité appelée méthode d'orthogonalisation est sélectionnée. Une description détaillée de la méthode retenue est fournie. L'objectif de ce chapitre est d'analyser l'avantage de sélectionner les paramètres en fonction de leur estimabilité par rapport à leur sélection en fonction de leur importance, et son effet sur l'ensemble de la méthodologie de calibrage. Les paramètres estimables sont les paramètres les plus importants et les plus identifiables : ils ont le moins d'interaction.

Les critères utilisés pour évaluer l'approche expérimentée sont l'amélioration de l'identifiabilité des paramètres après calibrage et la performance prédictive du modèle calibré. Ainsi, le premier indicateur utilisé est la distance de Janson Shannon. Il est utilisé pour calculer la distance entre les a priori et les a posteriori afin de quantifier l'identifiabilité des paramètres séparément. Le deuxième indicateur (ID) est proposé spécifiquement pour cette étude. Il calcule l'identifiabilité totale du modèle. Le troisième indicateur est l'erreur quadratique moyenne (RMSE) entre les résultats du modèle calibré et les mesures virtuelles. De plus, le DIC (deviance information criteria) qui est basé sur la fonction de vraisemblance est également utilisé pour évaluer la précision du modèle calibré. L'étude de cas présentée dans le chapitre précédent est conservée ici avec de légères différences. Au lieu de considérer six scénarios différents, seuls deux scénarios (2 et 3) correspondant à l'évolution libre sont considérés : le scénario 2 étant le scénario d'entraînement et le scénario 3 le scénario de test. La méthode de Morris est appliquée en premier, et les 40 paramètres les plus influents sont sélectionnés. Dans un second temps, la méthode Sobol est utilisée pour classer précisément ces 40 paramètres. Cela a nécessité un coût de calcul de 168 000 simulations.

Pour l'étude de cas considérée, il est montré que le classement des paramètres à l'aide de la méthode d'orthogonalisation est plus approprié que de les classer sur la base d'une analyse de sensibilité uniquement dans le cas où peu de paramètres doivent être calibrés. Une interaction significative peut exister entre les plus influents, qui peuvent être identifiés et pris en compte par l'analyse d'identifiabilité. Si plus de paramètres sont inclus, les deux méthodes donnent des résultats similaires.

L'effet du nombre de paramètres est également traité dans ce chapitre. L'étude de cas du chapitre 2 est conservée avec le classement de sensibilité correspondant. La méthode consiste à effectuer un calibrage avec un nombre croissant de paramètres, en commençant par calibrer uniquement le plus influent jusqu'à calibrer les 15 premiers paramètres. Le calibrage étant stochastique, il est répété dix fois pour chaque ensemble de paramètres. L'indicateur de Janson Shannon, RMSE, et le DIC sont également utilisés comme indicateurs dans cette analyse. Le but de cette analyse est d'évaluer le comportement du calibrage en termes de performance prédictive du modèle d'une part et d'identifiabilité des paramètres d'autre part avec un nombre différent de paramètres. Pour l'étude de cas considérée, on constate que le calibrage des trois premiers paramètres est le plus précis, cependant, le calibrage des huit premiers paramètres a donné une précision presque similaire, même si les six premiers paramètres s'avèrent plus identifiables que les autres.

4.1 Introduction

The importance of calibrating a subset of the building energy model parameters was clarified in the previous chapters. In chapter 2, the calibration methods are assessed following a sensitivity analysis which ranks the parameters in terms of importance. The set of most important parameters that are usually considered for calibration might not be the most estimable parameters: significant interaction might exist within the most influential parameters, which makes identification more difficult. In addition, identifiability problems may arise due to non-informative or insufficient data. Identifiability analysis allows to select the parameters that are most influential and have the least degree of interaction. The effect of the data on identifiability is not considered in this work.

In this chapter, based on literature, an identifiability method is selected. A detailed description of the retained method is provided. The aim of this chapter is to analyse the advantage of selecting the parameters based on their estimability compared to selecting them based on their importance, and its effect on the whole calibration methodology. Estimable parameters are the parameters that are the most important and most identifiable: they have the least degree of interaction. To this end, different criteria are applied on the calibrated posteriors as explained in details in section 4.2.2.

Several aspects concerning Bayesian inference applied to BEM require further analysis. One of the issues with calibration is the number of parameters to be estimated. On the one hand, including too many parameters lead to un-identifiability problems due to interaction; on the other hand, calibrating only few parameters is necessitates setting some of the important parameters at fixed values, which is subjected to uncertainties and error and could influence the estimation of the other parameters. Normally, neither too many nor too few parameters are considered. However, in the literature there is no clear answer to how many parameters should be included. Moreover, there is no sufficient analysis on this topic based upon different case studies, which would be needed to draw a conclusion on the recommended number of parameters allowing to achieve good calibration practice. In the second part of this chapter, the focus is oriented towards this issue and the effect of the number of parameters on the calibration performance is analysed.

4.2 Identifiability analysis

A brief introduction to the identifiability analysis, its importance, methods, and applications is provided in chapter 1. In this section, the selected identifiability method is defined and explained briefly, then it is applied to a case study.

The terms structural identifiability and sensitivity-based identifiability methods are used interchangeably in this thesis even though they could be separated. The main difference is that structural identifiability methods directly use the model structure to perform the identifiability analysis (Miao et al. 2011). Due to this, many structural methods only apply to linear (e.g. Laplace transforms), or simple non-linear models (power series expansion, similarity transform, direct test, etc.). Other structural identifiability methods apply to general non-linear models such as the differential algebra methods. However, the application of these methods to large systems such as BEMs could be quite complex. There are some open source software that apply these concepts such as DAISY (differential algebra for identifiability of systems (Bellu et al. 2007), however they are prohibitive for large systems (Rouchier 2018). Accordingly, the structural methods are not retained in this study.

Sensitivity-based methods do not directly use the structure of the model, however they benefit from the sensitivity matrix to perform the identifiability analysis, which makes them simple to implement. Thus, the methods selected in this chapter belong to this family.

The collinearity index method could be misleading since it does not account for the importance of the parameters. With a combinatorial analysis it computes the collinearity indices of different possible combinations of different sizes no matter how important the parameters are. It also requires the definition of a collinearity index threshold to decide whether the subset of parameters is identifiable or not. A recent study in the field of solids and structures (Zhang et al. 2022) applied this method and confirmed that there is a risk to discard highly influential identifiable parameters because, in their case, the collinearity index of these parameters is larger than the recommended threshold in literature which is between 15 and 20. Accordingly, this method is not retained in this chapter.

The correlation method suffers from two main drawback; the first is that it computes the correlation between pairs of parameters. Therefore, if significant correlation exists between a set of more than two parameters, the correlation method will not detect it (Quaiser and

Mönnigmann 2009). The second drawback is that it only accounts for correlations and it does not account for the importance of each parameter separately.

Quaiser and Mönnigmann (2009) compared four different sensitivity-based methods (correlation method, eigenvalue method, orthogonalisation method, and PCA method) with three different models. They concluded that for the three models, the eigenvalue and the orthogonalisation methods overcome the rest . In this chapter, the aim is not to compare different identifiability methods, however, it is to analyse the effect of a validated identifiability method on the calibration process. Accordingly, since the eigenvalue and the orthogonalisation methods showed similar performance in literature, only one out of the two methods is selected that is the orthogonalisation method.

The main objective of the study is to assess the behaviour of the calibration methods with different ways of parameters selection and to analyse how the identifiability analysis retained in this chapter enhances the global identifiability of the calibration method. Accordingly, the identifiability of the estimated posteriors is checked after calibration using the different criteria presented in section 4.2.2.

In this section, the orthogonalisation method selected among different sensitivity-based methods is firstly presented and explained in details, then the methodology and criteria used to assess the performance of the method are elaborated.

4.2.1 Orthogonalisation method

The orthogonalisation method was originally proposed by Yao et al (2003). The idea is that it re-ranks the parameters from the most to the least estimable. Most estimable means that the parameters are important and at the same time, the combination between the parameters has the least degree of interaction.

The procedure is to firstly compute the sensitivity indices matrix. This matrix should comprise k columns, each corresponding to a separate parameter and t_N rows, each corresponding to a given time step. Each column represents a vector of sensitivity indices of one parameter for each time step. The sensitivity matrix index is defined as follows:

$$S = \begin{pmatrix} s_{t_1,1} & s_{t_1,2} & \cdots & s_{t_1,k} \\ s_{t_2,1} & s_{t_2,2} & & s_{t_2,k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{t_N,1} & s_{t_N,2} & & s_{t_N,k} \end{pmatrix} \quad (4.1)$$

The most important parameter is selected as the most estimable parameter. Then, all the columns of the sensitivity matrix are projected onto the chosen column. This projection forms another matrix S_p having similar dimensions as the sensitivity matrix S.

$$S_p = S_L (S_L^T S_L)^{-1} S_L^T S \quad (4.2)$$

where S_L is a $t_N \times L$ matrix with L being the number of estimable parameters selected. At the first iteration of the algorithm, L is 1 since S_L comprise only the most influential parameter. After each iteration, a new parameter is selected to be the next most estimable parameter and is then concatenated to S_L. At the end, S_L will contain the sensitivity vectors in the original sensitivity matrix S but in a different order: the first column is now the sensitivity vector of the most estimable parameter and the last one is that of the least estimable one.

The selection of the next most estimable parameter at each iteration is done via the residual matrix R_L which is computed by subtracting S_p from the original sensitivity matrix S. The column having the largest magnitude in R_L is then selected as the second most estimable parameter. The residual matrix accounts not only for interaction, but also for the level of importance of each parameter: if two parameters are correlated, the method selects the one which is more influential. The steps are summarised as follows (adapted from Yao et al., 2003):

Orthogonalisation method

1. Calculate the magnitude of each column of the sensitivity matrix *S*.
2. Select the parameter whose column in *S* has the largest magnitude as the first estimable parameter.
3. Mark the corresponding column as S_L (L=1 for the first iteration).
4. Project the columns in S onto S_L : $S_p = S_L\ (S_L^T S_L)^{-1} S_L^T S$.
5. Calculate the residual matrix $R_L = S - S_p$.
6. Calculate the sum of squares of the residuals in each column of R_L. The column with the largest magnitude corresponds to the next estimable parameter.
7. Select the corresponding column in *S*, and augment the matrix S_L by including the new column.
8. Advance the iteration counter by one and repeat steps 4 to 7 until the column with the largest magnitude in the residual matrix is smaller than a prescribed cut-off value.

4.2.2 Methodology and criteria

To assess the importance of identifiability analysis prior to calibration, the most influential, then the most estimable parameters are calibrated. Since, the calibration is stochastic and different results might be obtained with different runs, the calibration is repeated several times for each model. This overcomes the stochasticity of calibration and makes the comparison more reliable. Since this methodology can be computationally intensive, the minimum threshold of RMSE is set to 0.05℃. This allows the calibration to converge faster. Here, the model refers to the BEM with the selected parameters for calibration: choosing the most important parameters is a model, and choosing the most estimable is another model.

Virtual data based on known parameters values are generated. This allows analysing how well each parameter is estimated and how close it is to its true value, which is not attainable with real measurements. At each repetition of calibration, the parameters that are not included in calibration are retained to their true values.

Since there is no clear recommendation on the maximum number of parameters that can be included for calibration, an increasing set of parameters is considered. That is firstly, the two

most important parameters are calibrated and then the three most important ones are calibrated and so on. The same is done for the most estimable parameters. The aim here is to compare the results attained after calibrating the most influential parameters against calibrating the most estimable parameters. To undergo this comparison, different criteria are retained.

The Jensen-Shannon *JS* distance is a criterion used to measure the distance between the prior and the posterior. The larger the distance, the more identifiable the parameter is. *JS* is built on the *KL*-divergence between two distributions as follows:

$$JS(P||Q) = \sqrt{\frac{KL\left(P||\frac{P+Q}{2}\right) + KL\left(Q||\frac{P+Q}{2}\right)}{2}} \tag{4.3}$$

where P and Q represent the two distributions between which the similarity is computed. KL-divergence is not symmetric and thus it is not considered to be a distance, however, JS could be considered as a symmetric version of the KL-divergence:

$$KL(P||Q) = \int P(\theta) \times \log\frac{P(\theta)}{Q(\theta)} d\theta = \sum_{i=1}^{N} P(\theta_i) \times \log\frac{P(\theta_i)}{Q(\theta_i)} \tag{4.4}$$

This metric should be used carefully. Even if the posterior is very close to the prior, one cannot conclude the presence of un-identifiability. This could be related to the proper selection of the prior. In our case, no conclusions concerning the identifiability issues will be drawn from the *JS* distances of one model, however, a comparison on the *JS* distances of the two models is considered.

Another misleading result should also be accounted for. A parameter could be perturbed linearly with another parameter by maintaining approximately similar results if a linear correlation exists between them. Thus, even if the posteriors of those two parameters are different from their priors, this does not mean that they are identifiable. To avoid this, the JS distance is used to measure how narrow the posterior distribution is compared to the prior given that both have the same mean value. Since a virtual data is considered and the true values of the parameters are known, the priors are centred on the true values. This allows the posteriors to have same means as those of the priors, which means that the JS distance between the prior and the posterior represents how much the variance around the true value is reduced. This is illustrated in Figure 4.1, where the distribution in case 2 is narrower than that in case 1, and this

is explained by the larger value of JS. The interpretation of JS values should not be confused by cases 3 and 4 since the posteriors in the cases studied do not shift away from the true values (priors means). Biased priors are not used in this study since it will be computationally more intensive especially that calibration is repeated several times. Moreover, the considered method with the chosen priors answers the question of identifiability that is looked for in this section. This criterion enables one to estimate the identifiability of each parameter separately.

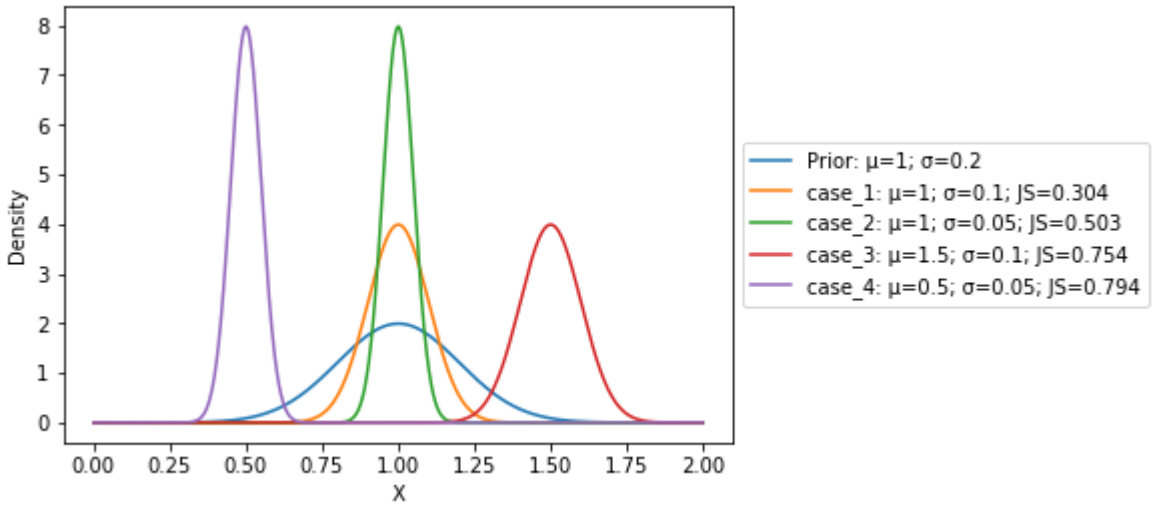

Figure 4.1: Illustration of JS criterion

To compute the total identifiability of the model, another indicator is proposed. Through calibration, the samples with low likelihood or high RMSE in the context of ABC are discarded in favour of those that yield higher likelihood or lower RMSE. This is done iteratively until the posterior is reached. If the set of parameters are unidentifiable, then the samples retained for the posterior might be just sets of different possible combinations that respect the convergence criterion of the sampler. Therefore, if posteriors are constructed from the average and variance of these samples and then those posteriors are propagated, the resulting likelihoods or RMSE will be worse than those achieved by the samples drawn from the MCMC sampler.

This is clearly observed in ABC algorithms, where even if the minimum threshold of RMSE identified for convergence is reached, the propagation of the posteriors may yield higher values of RMSE. The difference between the minimum threshold δ_{min} and the average over all RMSEs resulting from the posteriors propagation is thus considered as an indicator of the model total identifiability.

$$ID = \frac{1}{N}\sum_{i=1}^{N} RMSE(\eta(\theta_i), Z) - \delta_{min} \tag{4.5}$$

Another criterion assesses the model predictive performance. That is how much the model fits to the data. Since for all the models, the same data is used, it is possible to use a model selection criterion. AIC (Akaike Information Criterion) is a well-known and widely used model selection criteria (Akaike 1974). This metric is calculated from the maximum likelihood estimate of the parameters, which makes it more suitable for frequentist approaches such as profile likelihood. In this thesis, there is an interest not only to quantify the best parameters values that fit to data, but to generate a distribution of values that describe the uncertainty of the estimates. Thus, taking only the maximum likelihood in AIC (equation 4.6) discards the additional information that can be obtained from the posterior PDF.

$$AIC = -2\ \log p(Z|\theta_{MLE}) + 2K \tag{4.6}$$

In equation (4.6), K is the number of parameters in the model, and θ_{MLE} is the parameters values which corresponds to the maximum likelihood. To this end, DIC (deviance information criterion) is more suitable in the Bayesian context and is retained. The idea is that instead of computing the log-likelihood at θ_{MLE}, it is calculated at the posterior mean. Moreover, the second term in AIC is replaced with a data-based bias correction p given as follows (Gelman et al. 2013):

$$I = 2\left(\log p(Z|\bar{\theta}) - \frac{1}{N}\sum_{n=1}^{N} \log p(Z|\theta_n)\right) \tag{4.7}$$

where $\bar{\theta}$ is the mean of the parameters posteriors. The second term in I computes the log-likelihood of all the posteriors samples and averages them. This shows how DIC accounts for the posterior against the AIC, which is more suitable for point estimate problems. The DIC is then computed as follows:

$$DIC = -2\ \log p(Z|\bar{\theta}) + 2I \tag{4.8}$$

The value of the DIC has no interpretable meaning if considered alone, instead, the comparison between two values gives information about which model is better in terms of fitting to data than the other. The lower the value of DIC, the more the model fits to the data.

To sum up, the criteria used in this analysis are the JS, and ID as identifiability indicators and DIC as model predictive performance indicator.

4.2.3 Case study

The case study presented in the previous chapter is retained here with slight differences. Instead of considering six different scenarios, only two scenarios (2 and 3) that correspond to the free evolution are considered: scenario 2 being the training scenario, and scenario 3 the test scenario. Compared to the previous studies, no heating power, ventilation, or internal gains are included.

In the following section, the aim is to assess the behaviour of calibration when estimating the most important versus the most estimable parameters. Morris' method was found to be robust and accurate in estimating the parameters ranking as shown in chapter 2, but still there exists some influential parameters that are ranked differently from Sobol method. Even if the difference between Sobol and Morris methods is small, this would bias the comparison, since a sequential selection of the parameters is taken into account as described in section 4.2.2 and not the cluster of the most important parameters. This means that if there is no accurate ranking, the selected parameters may not be the true most important ones. Accordingly, Sobol (reference method) is retained to rank the parameters in terms of importance, then identifiability analysis is performed on the Sobol indices.

However, due to its computation cost, and since Morris ranks all the parameters with sufficient accuracy, Morris method is used as a first step to screen the less influential parameters, then Sobol is applied to the 40 most influential parameters ranked by Morris method. This decreases a lot the computational cost of Sobol method since it is highly dependent in the number of parameters: instead of 460,000 simulations, only 168,000 simulations are required.

4.2.4 Results and discussion

4.2.4.1 Sensitivity analysis

Sobol method is applied to the given case study on the second scenario. Figure 4.2 shows the ranking of the first 8 parameters: specific heat of concrete wall $(c_{p,concW})$, conductivity of

polystyrene wallmate (λ_{polW}), specific heat of concrete screed $(c_{p,concS})$, specific heat of reinforced concrete $(c_{p,concR})$, thickness of concrete wall (t_{concW}), conductivity of polystyrene styrofoam (λ_{polS}), specific heat of slab joist $(c_{p,J})$, and thickness of polystyrene wallmate (t_{polW}). Since the considered scenario resembles the free evolution of the building and since the building is made of concrete construction (external walls with concrete, floor with reinforced concrete and concrete screed), the specific heats of concrete, reinforced concrete, and concrete screed are classified within the four most influential parameters. The specific heat of concrete wall is the most influential since it is used in the external walls where it is subjected to weather variations more than the others, moreover, it is present in the four external walls compared to the presence of the rest only in the ground floor. The building is well insulated from the ground with polystyrene styrofoam, and from the walls with polystyrene wallmate. This explains why the conductivity of these two materials is also classified within the first most influential parameters.

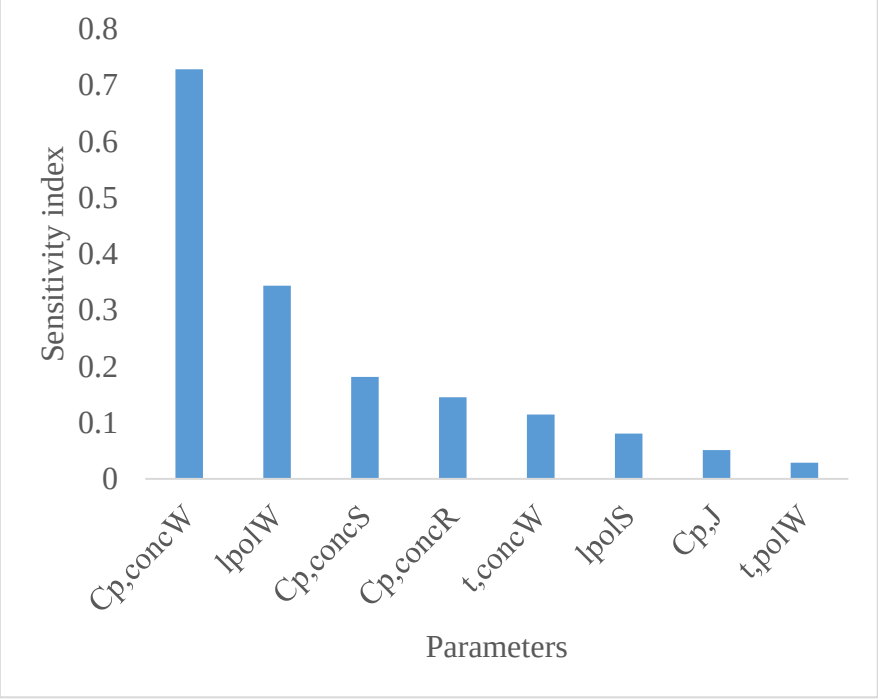

Figure 4.2: Sobol rank of the first 8 parameters

4.2.4.2 Estimability ranking

The orthogonalisation method is applied to the sensitivity matrix computed, and the parameters are re-ranked as shown in Table 4.1. It also shows the value of the threshold in the orthogonalisation method. This means that if the threshold is set at 0.16, the orthogonalisation

method states that only two parameters ($c_{p,concW}$, and λ_{polW}) can be estimated and the rest should not be included in parameter estimation. With a threshold of 0.04, the orthogonalisation method states that in addition to the first two, the specific heat of the reinforced concrete is the third most estimable parameter. Only six parameters are illustrated. The reason is that it is found after calibrating the first three parameters that more parameters are significantly less identifiable, and that both rankings start to yield equal results. Moreover, this method is computationally intensive as it requires multiple calibration run, so it is decided to stop at 6 parameters.

Table 4.1: Sobol and estimability ranks

Rank	Sobol method	Orthogonalisation	Threshold
1	Specific heat of concrete wall $(c_{p,concW})$	Specific heat of concrete wall $(c_{p,concW})$	
2	Conductivity of polystyrene wallmate (λ_{polW})	Conductivity of polystyrene wallmate (λ_{polW})	0.16
3	Specific heat of concrete screed $(c_{p,concS})$	Specific heat of reinforced concrete $(c_{p,concR})$	0.04
4	Specific heat of reinforced concrete $(c_{p,concR})$	Conductivity of polystyrene styrofoam (λ_{polS})	0.0118
5	Thickness of concrete wall t_{concW}	Specific heat of concrete screed $(c_{p,concS})$	0.011
6	Conductivity of polystyrene styrofoam (λ_{polS})	Thermal bridge living room to exterior (ψ)	0.00315

The value of the threshold is important since, the number estimable parameters is based on it. In this section, the aim is to compare the two methods of selecting the parameters for calibration, and after that to analyse the effect of the threshold value. Accordingly, no cut-off value for the threshold is considered here: the most estimable parameters are selected sequentially as described in the section 4.2.2.

4.2.4.3 Identifiability assessment

The next step is to assess the performance of calibration based on the most important and most estimable parameters. For this reason, APMC (adaptive population Monte Carlo) algorithm, which is detailed in chapter 2, is retained since it showed the best performance against the other methods. The minimum threshold at which the algorithm converges to the posterior is taken to be RMSE = 0.01℃.

At each calibration repetition, the discarded parameters are fixed at their true values. Since sensitivity and identifiability methods rank the first two parameters equally, the calibration is applied starting from three parameters. APMC is run 20 times on the three most influential parameters ($c_{p,concW,}$ λ_{polW}, and $c_{p,concS}$), and then run again for 20 times on the three most estimable parameters ($c_{p,concW}$, λ_{polW}, and $c_{p,concR}$), and so on.

For all the runs, the posteriors are centred on the true values of the parameters. The difference between one run and the other is just related to how spread the posterior is. Accordingly, the Jensen Shannon distance is a good indicator in this situation to describe how well the parameters are identified from the data. For the rest of this section, calibrating the most important parameters is called I-model, and calibrating the most estimable parameters is called E-model. Figure 4.4 shows the distances between the priors and the posteriors of $c_{p,concS}$ and $c_{p,concR}$, which are the parameters that differ between the two models, in a form of boxplot. A boxplot combines five summary statistics that describe the data in one chart. The box shows the region where the interquartile range of the data exists: the data above first quartile (25 % centile) and below the third quartile (75 % centile). The horizontal line in the box represents the average value. The upper and lower ticks also called whiskers of the chart represent the minimum and maximum values in the data. It is obviously depicted that the difference between the prior and the posterior of $c_{p,concR}$ is much larger than that of $c_{p,concS}$. Given that both posteriors are centred on their true values as depicted in Figure 4.3, this means that the posterior of $c_{p,concR}$ is much narrower than the posterior of $c_{p,concS}$ which means that it is more identifiable, even though it is estimated by Sobol method that $c_{p,concS}$ is more important than $c_{p,concR}$. Note that the first two parameters are ranked equally by the two methods, so the results of calibrating only those two parameters are not shown.

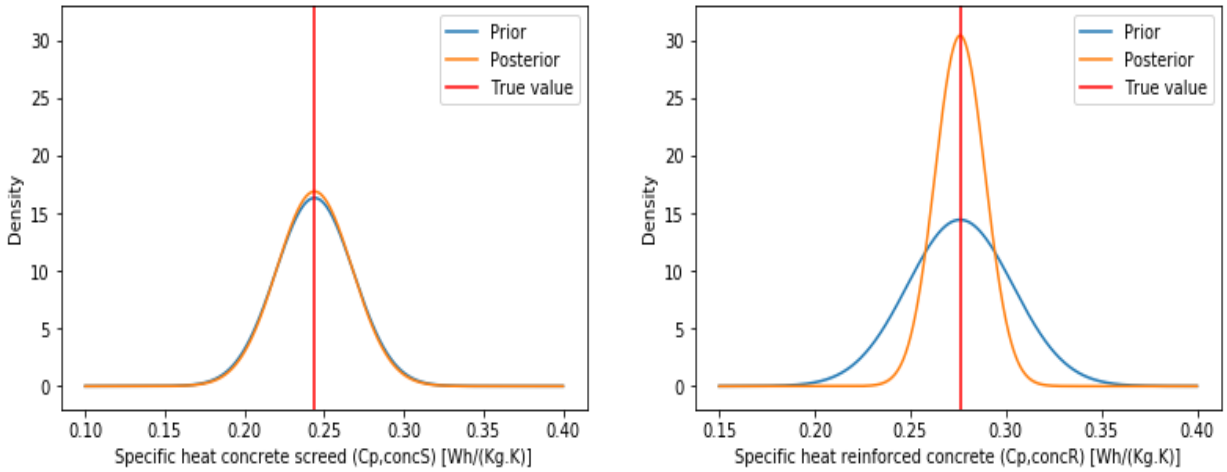

Figure 4.3: Posterior vs prior of Cp,concS and Cp,concR in set of three parameters

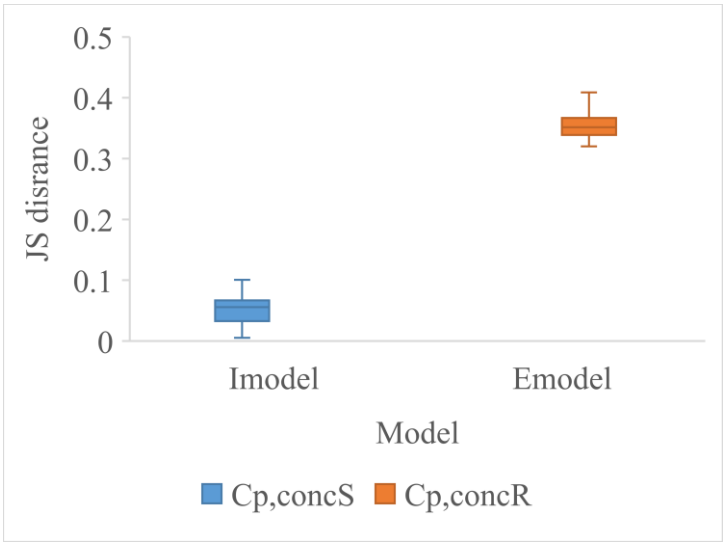

Figure 4.4: JS distance in both models for Cp,concS and Cp,concR (set of three parameters)

To validate the importance of $c_{p,concS}$ and $c_{p,concR}$, each one is calibrated alone without any other parameter. The two parameters can be correctly identified as depicted in Figure 4.5. Both parameters are similarly identifiable with a slight tendency for the $c_{p,concS}$ to be better estimated.

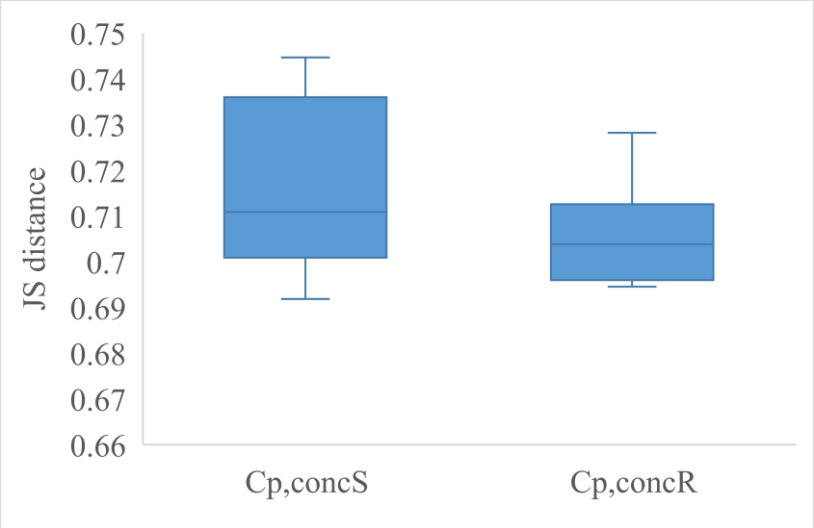

Figure 4.5: JS distance of parameter Cp,concS, and Cp,concR when estimated separately

This validates the ranking of the Sobol method, and thus, it could be said that the reason behind not being able to well identify $c_{p,concS}$ in the Imodel is the interaction with other parameters. To investigate this even more, the *JS* distances of $c_{p,concW}$, and λ_{polW} are computed for the two models as depicted in Figure 4.6. It shows that the *JS* distances of both parameters in the E-model are greater than those in the I-model, which means that $c_{p,concW}$, and λ_{polW} in the E-model can be identified better than in the I-model.

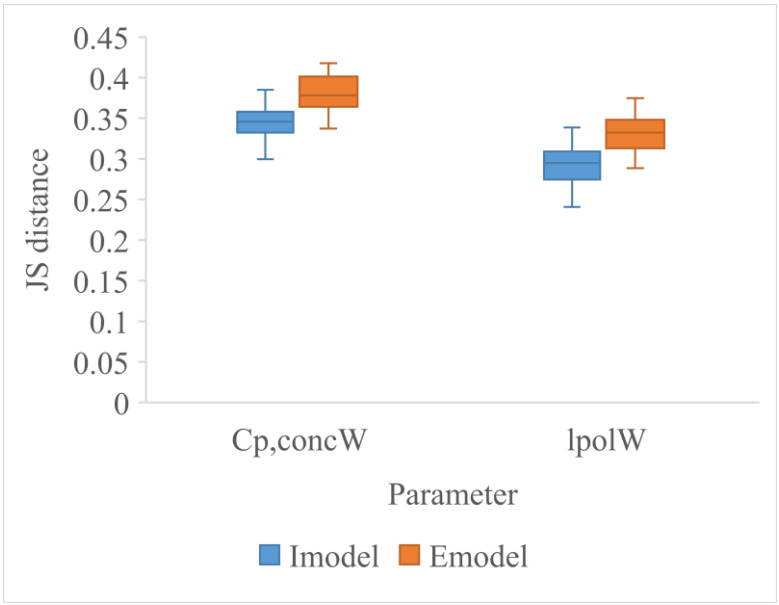

Figure 4.6: JS distance in both models for Cp,concW, and λpolW (set of three parameters)

To complement this analysis, ID and DIC are applied to both models (Figure 4.7). The plot on the left shows the DIC for each model. The whiskers of each model do not intersect which means that the minimum DIC obtained with the I-model is greater than the maximum

value obtained with the E-model. This shows that the E-model performs better in terms of fitting to the data than the I-model, which is consistent with the results, attained from assessing the identifiability of each parameter.

The plot on the right in Figure 4.7 shows the ID of both models for the set of three parameters (Table 4.1). The greater the value of ID, the larger the difference between the minimum threshold and the RMSE of the posteriors propagation as explained in section 4.2.2. Both models show that there is a certain degree of un-identifiability, where the average of the I-model and the E-model results are 0.09℃ and 0.07℃ respectively. For a perfectly identifiable model, this value should be zero. The E-model is less than that corresponding to the I-model, which means that E-model is more identifiable than the I-model. The RMSE of the propagation of the I-model and E-model can be extracted from the ID by adding 0.01 ℃ (the minimum threshold used in calibration) to the values of ID. This is done for the average ID values, and yields an RMSE of 0.1 ℃ and 0.08 ℃ for the I-model and E-model respectively, which is a 20 % difference. This difference is not negligible and thus, it can be said that E-model is better fitting to the data than I-model.

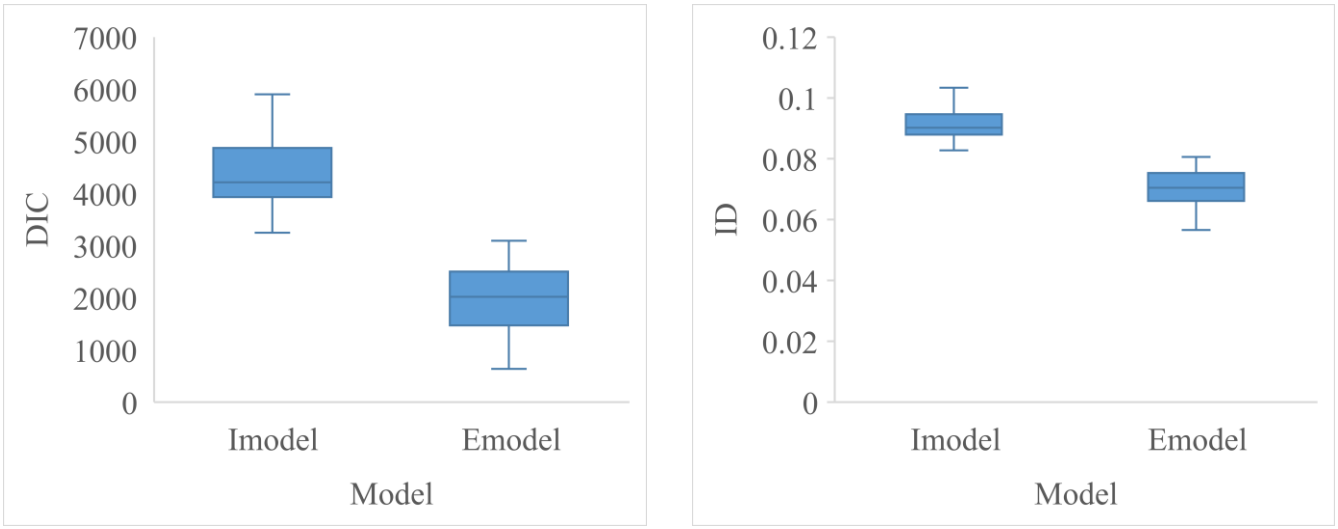

Figure 4.7: DIC and ID of both models for the set of three parameters

The identifiability of the two most estimable and important parameters ($c_{p,concW}$, and λ_{polW}) in each parameter set, starting from three ending with six parameters, is depicted in Figure 4.8. With four parameters, the additional parameter in the E-model is the thermal conductivity of polystyrene Styrofoam, λ_{polS}, which is added to the ground floor of the house. This parameter possesses some degree of interaction with λ_{polW} which explains why the degree of identifiability of λ_{polW} decreases in the E-model after including λ_{polS}. It is also evident that

λ_{polS} does not have significant interaction with $c_{p,concW}$, which explains why the identifiability of the latter remained almost the same from set three to set four.

In the I-model, the identifiability of λ_{polW} did not significantly vary since the additional parameter at set four is $c_{p,concR}$, which does not have a significant degree of interaction with λ_{polW}. This also explains why the variation in the identifiability of $c_{p,concW}$ is not significant, which is consistant with the results of the orthogonalisation method that states that the degree of interaction between these two parameters is small.

Having said that, the E-model performs worse than the I-model in identifying the conductivity of polystyrene wallmate. It is still better than the I-model in identifying the specific heat of concrete wall since with this set of parameters the specific heat of the concrete screed which is more correlated to the conductivity of the polystyrene wallmate, is not calibrated.

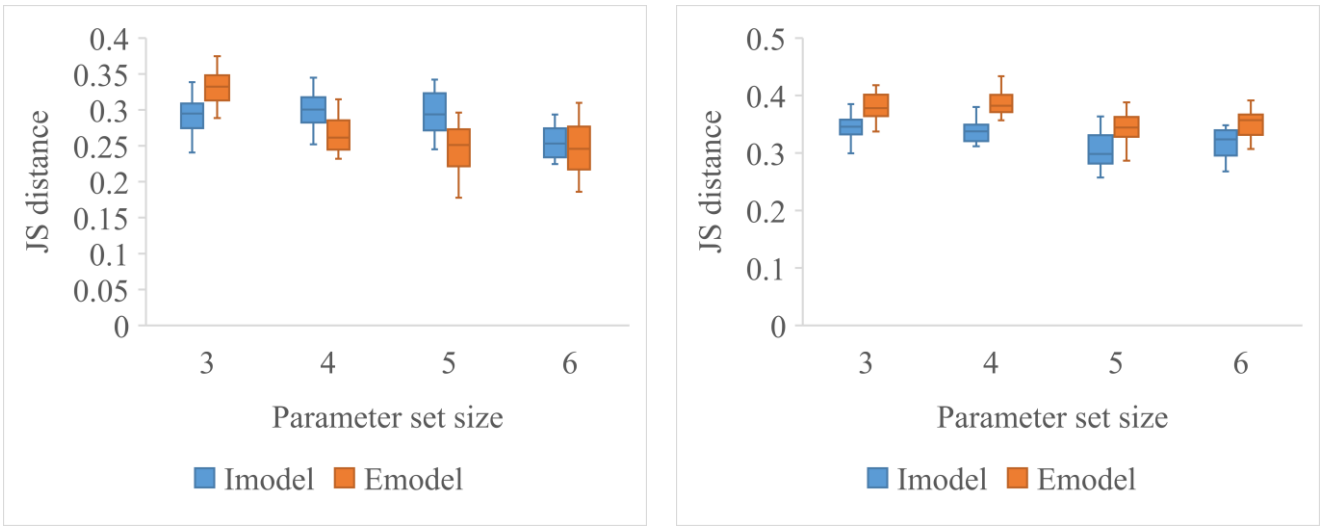

Figure 4.8: JS distance of λ_{polW} (left), and $c_{p,concW}$ (right) with all sets in both models

The additional parameter at set five in the I-model is t_{concW}. This causes an additional degree of un-identifiability on the estimation of $c_{p,concW}$. This also explains why the identifiability of λ_{polW} did not show significant variation, since there is no direct interaction between these two parameters.

On the contrary, the additional parameter at set five in the E-model is $c_{p,concS}$. Accordingly, the identifiability of $c_{p,concW}$ decreased, however, the conductivity of polystyrene wallmate identifiability is the same. At this set, the E-model performs better than the I-model in identifying $c_{p,concW}$ since in the Imodel, there exists two parameters that seem to interact

with $c_{p,concW}$ that are: $c_{p,concS}$ and t_{concW}, however, in the Emodel, t_{concW} is still not accounted for in calibration. On the contrary, the I-model performed better in identifying λ_{polW} than the Emodel, since λ_{polS} is still not present in the Imodel.

In set six, the I-model includes λ_{polS}. Consequently, the identifiability of λ_{polW} decreased as depicted in Figure 4.8 compared to an approximate similar identifiability for $c_{p,concW}$. In the Emodel, the additional parameter is the thermal bridge between the living room and the exterior and it caused no significant variation on $c_{p,concW}$ and λ_{polW}. Similar performance is obtained between both models in estimating these two parameters with a slight advantage for E-model to estimate $c_{p,concW}$.

The average of the JS distance is estimated for the parameters. This average value represents the degree of identifiability in the posteriors. This is done for each calibration run. Figure 4.9 shows the average values as boxplots for the two models with an increasing number of parameters. It shows that the largest difference between the two models is with three parameters, and the difference becomes smaller with an increasing number of parameters. With four parameters, there is a slight difference between both models, where the median of the E-model is greater than that of the I-model but the whiskers are close.

This graph only gives a representative estimation about the difference between the two models. If two models have the same average value of the JS distance, it does not mean that both models will perform equally in fitting to the data. For example, if JS of an influential parameter is 1 and of an un-influential parameter is 0, then the average is 0.5, which is the same exact average if the JS of the two parameters are switched. Thus, these results need to be complemented by looking at the identifiability of the parameters separately on the one hand, and applying other indicators such as ID and DIC on the other hand.

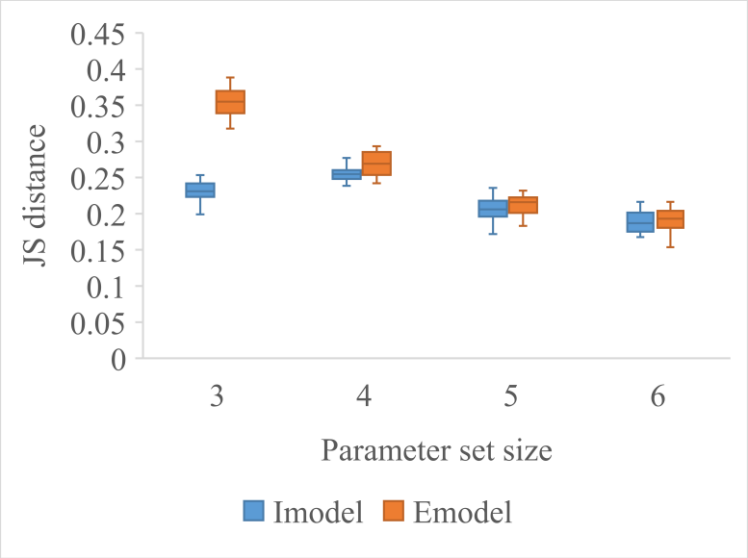

Figure 4.9: Average over JS distances for each parameter set size

Figure 4.10 shows the identifiability of each of the four parameters in set four in both models. It could be seen that the difference in the identifiability of $c_{p,concR}$ between the two models is not very significant. λ_{polS} is estimated slightly better than $c_{p,concS}$, eventhough it is less influential. This reflects the importance of studying identifiability between the parameters prior to calibration. Since, the difference in the estimation of $c_{p,concW}$ in favour of E-model (E-model estimates it better) is more significant than the difference in the estimation of λ_{polW} in favour of I-model (I-model estimates it better), it could be said that the E-model would fit better to data than the I-model.

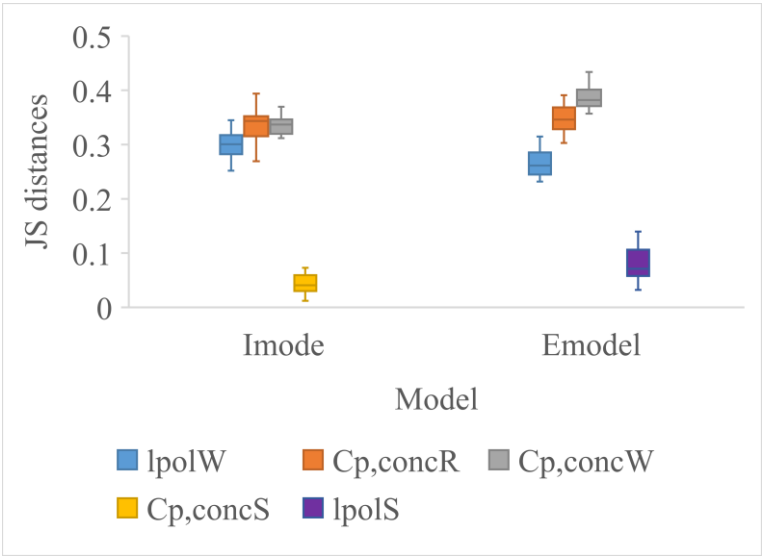

Figure 4.10: JS distance of all the parameters in the set of four parameters

This is confirmed in Figure 4.11. The plot on the right shows the DIC of each model and it states that the E-model fits better to data. The one on the left shows the ID of each model and it states that the E-model has a higher degree of identifiability than the I-model. This is consistent with the observations extracted from Figure 4.9. The RMSE are also extracted as previously. The RMSE of the I-model is 0.106℃ and 0.095℃ respectively, which represents a 10 % difference. The difference is not as significant as it was in the previous set.

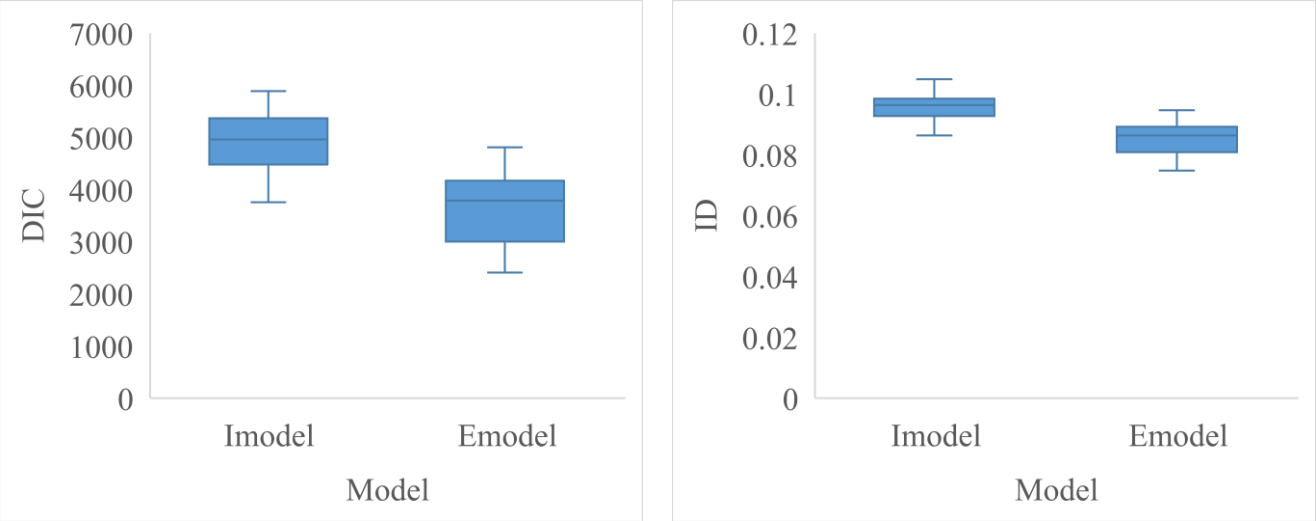

Figure 4.11: DIC and ID of both models for the set of four parameters

For the set of five parameters, Figure 4.12 show that both models have very similar performance in terms of ID and DIC with a slight better performance for I-model, since the interquartile regions significantly intersect and the whiskers are also close to each other. The better estimation of $c_{p,concW}$ in the Emodel is in this case compensated by a better estimation of λ_{polW} in I-model as depicted in Figure 4.13. The RMSE of the I-model and the E-model are 0.116℃ and 0.119℃ respectively, which represents a negligible difference of 2.5% in favour of I-model. These RMSE values in addition to the data found in the boxplot mean that both rankings at this point behave similarly.

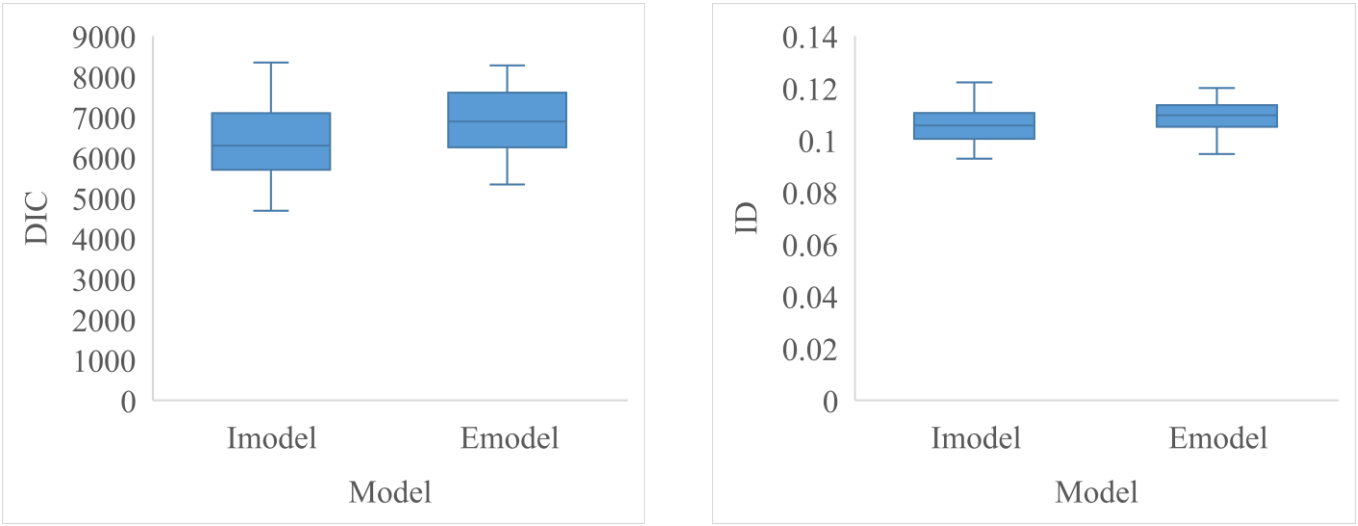

Figure 4.12: DIC and ID of both models for the set of five parameters

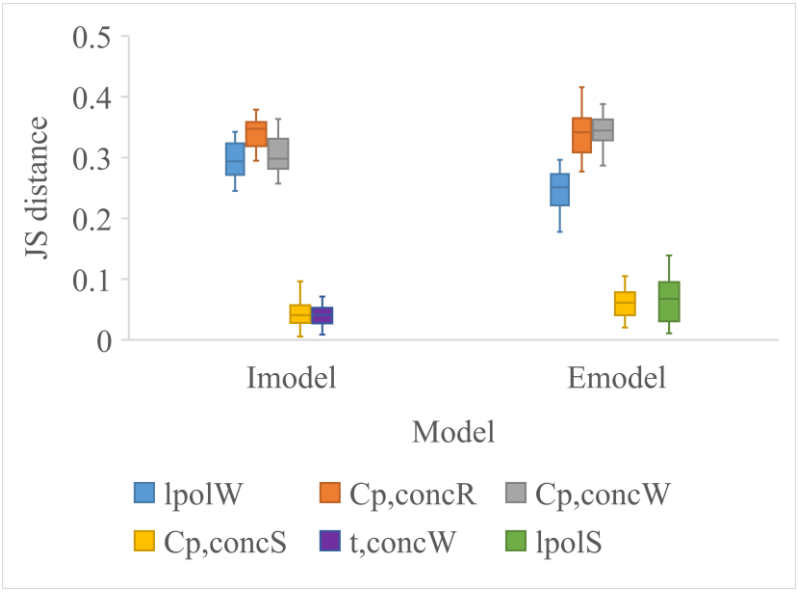

Figure 4.13: JS distance of all the parameters in the set of five parameter

For the last set of parameters (Figure 4.14), all the parameters have more or less similar degrees of identifiability except for $c_{p,concW}$ which is better estimated in the Emodel as discussed previously. Consequently, both models perform similarly in terms of ID and DIC with a slight advantage for the E-model as depicted in Figure 4.15. In terms of RMSE, the difference is 7 % on average in favour of the E-model. The reason behind this slight increase in the difference percentage between the set of five parameters and the set of six parameters is as discussed earlier, the addition of the thickness of concrete wall, which has direct interaction with its specific heat.

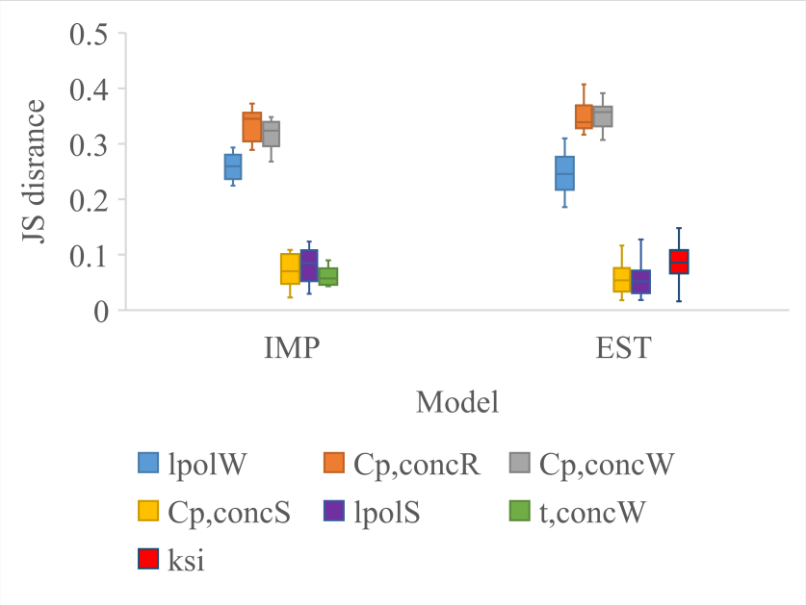

Figure 4.14: JS distance of all the parameters in the set of six parameter

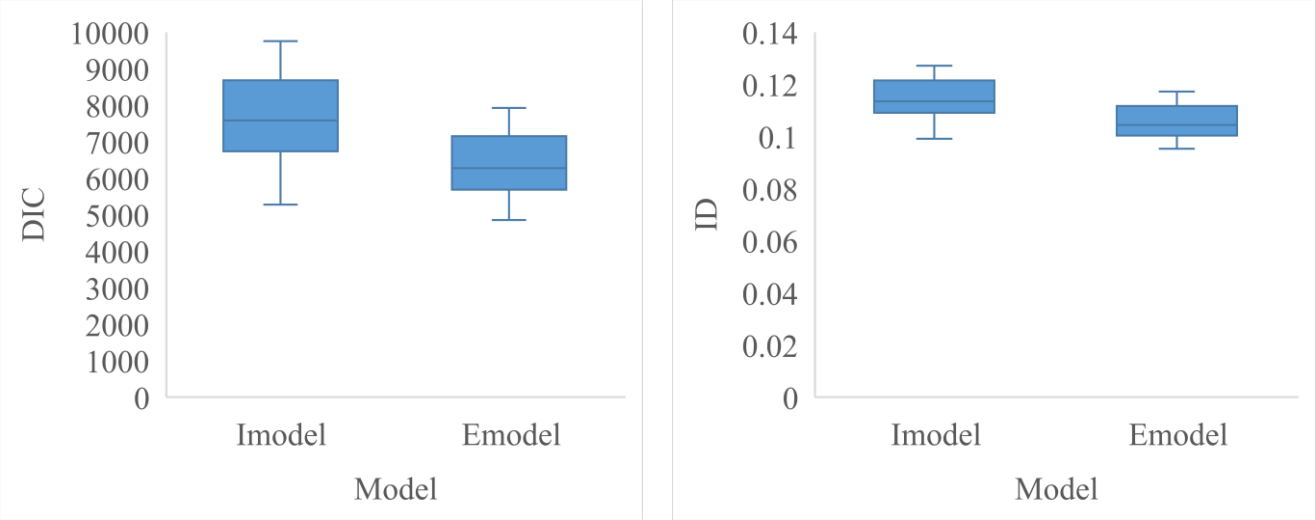

Figure 4.15: DIC and ID of both models for the set of six parameters

It is also worth noting that identifiability of the parameters other than the three most estimable is significantly lower than these three: the difference could be clearly visualised in Figure 4.10, Figure 4.13, and Figure 4.14. Referring back to Table 4.1, the estimability threshold corresponding to the first three estimable parameters is 0.04 and then decreases to 0.011 for the set of four parameters. Yao et al. (2003) used a value of 0.04 as a threshold below which the parameters are discarded. However, this value was selected arbitrarily, and it is not proven that this is the recommended value, since it might vary with different case studies. For this case study, it seems like a threshold of 0.04 is able to classify the parameters that can be easily estimated from data and the parameters that are more difficult to estimate.

4.2.5 Conclusion

In this section, an identifiability analysis based on orthogonalisation is implemented and applied to a virtual case study. Morris' method followed by Sobol's method is applied to screen out the less influential parameter and to ensure a correct rank of the most influential parameters. Orthogonalisation is then applied on the Sobol indices matrix.

A comparison is then established between calibrating the most important parameters ranked by Sobol method and the most estimable parameters re-ranked by the orthogonalisation method. To ensure the reliability of this comparison, and to diminish the effect of the calibration variability on the results, each calibration is repeated 20 times. Three different indicators are used: Jensen-Shannon distance between the priors and the posteriors to quantify the identifiability of each parameter, ID distance that is used to quantify the identifiability of the model regardless of each parameter separately, and DIC that estimates how the calibrated model fits to data.

Based on these indicators, it is shown that calibrating the parameters based on estimability rank is better than taking the most important ones. Significant interaction may exist between the most important ones, which can be identified and accounted for through the identifiability analysis. If more parameters are included, the estimation based on both rankings become similar: the interaction avoided between the first more estimable parameters appear again with more parameters included.

No cut-off value for the orthogonalisation method threshold is identified, since the aim is to compare both rankings. In fact, the orthogonalisation method terminates when a threshold value is reached, which means that the parameters that are not ranked are considered not estimable. In this case study, it is shown that the parameters that were more accurately estimated than the others are those ranked by the orthogonalisation method with a threshold of 0.04. This is consistent with the work of Yao et al. (2003), but further study is needed to confirm this value. It is very useful if a generalised value of this threshold could be recommended, since this helps the calibration practitioners to select the appropriate number of parameters for calibration.

Moreover, the identifiability analysis in this chapter is coupled with Sobol sensitivity method. This is important for analysing the effect of the identifiability analysis where an accurate ranking of the parameters is required. However, in practice, this could not be feasible due to the Sobol method computational burden. In this case, it is very important to assess the

performance of the identifiability analysis if coupled with Morris method. This is not considered in this thesis, but it is an important study to be performed in the future.

4.3 Effect of number of parameters

Few studies focused on the effect of the number of parameters on the identifiability of the calibration approach and its computational cost. Chong and Menberg (2018) studied the effect of calibrating two to six parameters. The study was conducted on real data and not on virtual data, so the un-identifiability indicator was the increase in the posterior uncertainty. They found that un-identifiability problems occurred starting from calibrating four parameters. Kang and Krarti (2016) investigated the influence of the number of parameters on the posteriors. They calibrated one to 11 parameters. The root mean square error, coefficient of variation, and mean bias error were used to indicate the dispersion and uncertainty in the posteriors. They found that the posteriors errors and uncertainties increase gradually with the increase of the number of parameters. They did not propose what the best number of parameters corresponding to their study is.

4.3.1 Methodology and criteria

The aim is to evaluate the effect of the parameters number on the parameters estimation on the one hand and on the model predictive performance on the other hand. To do that, an increasing number of parameters is considered. This means that, firstly, the most estimable parameter is calibrated, then the two most estimable are calibrated and so on.

The ranking based on estimability, which is attained after applying the orthogonalisation method is used instead of the importance ranking, since it showed better performance in the previous section. Moreover, applying it on another case study (described below) is essential to validate its performance on the one hand and to assess the best value of the cut-off threshold. In the previous section, it is shown that a cut-off threshold of 0.04 is valid to cluster the estimable parameters from the non-estimable ones. It is important to assess the generalisability of this value with different case studies.

The case study used in chapter 2 is retained. Accordingly, it is not needed to run again a sensitivity analysis on the parameters. The results of Sobol method are used as a basis for the identifiability method.

Since, the calibration is stochastic and different results might be obtained with different runs, the calibration is repeated 20 times for each set of parameters. This overcomes the stochasticity of calibration and makes the comparison more reliable.

Virtual data based on known values of the parameters are generated. This allows analysing how well each parameter is estimated and how close it is to its true values, which is not attainable with real measurements. At each repetition of calibration, the parameters that are not included in calibration are retained to their true values. To analyse the identifiability of each parameter, the Jensen-Shannon distance is used.

The deviance information criterion (DIC) is also retained in this section to quantify the ability of each calibrated model to fit well to the data. In order to account for the uncertainties in the discarded parameters and to be able to analyse the relation between estimating few parameters versus estimating more parameters on the model predictive performance, the uncertainties of the calibrated and un-calibrated parameters are propagated. Firstly, before running any calibration, the priors of all the parameters are propagated and the DIC value is recorded. Secondly, the first most estimable parameter is calibrated with all the other parameters set at their true value. This ensures that calibration is performed in controlled conditions: no noise or uncertainties are present which facilitates the analysis. Then the posterior of the most estimable parameter along with the priors of the rest are propagated and the DIC values are recorded. This is applied to all the parameters sequentially.

4.3.2 Results and discussion

The orthogonalisation method is applied to the sensitivity vectors of the Sobol method. Table 4.2 shows the results of the importance and estimability ranking with different cut-off values of the threshold.

Table 4.2: Sobol and estimability ranking

Rank	Sobol method	Orthogonalisation	Threshold
1	Ventilation flowrate $(\dot{V})$	Ventilation flowrate $(\dot{V})$	
2	Heating power (Q_p)	Specific heat of concrete wall $(c_{p,concW})$	0.898
3	Specific heat of concrete wall $(c_{p,concW})$	Heating power (Q_p)	0.427
4	Conductivity of polystyrene Wallmate (λ_{polW})	Albedo (Alb)	0.304
5	Internal gains (Q_d)	Conductivity of polystyrene Wallmate (λ_{polW})	0.196
6	Albedo (Alb)	Internal gains (Q_d)	0.04
7	Specific heat of concrete screed $(c_{p,concS})$	Conductivity of polystyrene Styrofoam (λ_{polS})	0.015
8	Conductivity of polystyrene Styrofoam (λ_{polS})	Window heat transfer coefficient (U_w)	0.0087
9	Thickness of concrete wall t_{concW}	Specific heat of concrete screed $(c_{p,concS})$	0.007327
10	Thermal bridge living room to exterior (ψ)	Thickness of concrete wall t_{concW}	0.0066
11	Window heat transfer coefficient (U_w)	Specific heat of reinforced concrete $(c_{p,concR})$	0.006
12	Thickness of polystyrene Wallmate $(t_{,polW})$	Infiltration flowrate (inf)	0.00265
13	Specific heat of reinforced concrete $(c_{p,concR})$	Thermal bridge living room to exterior (ψ)	0.0016
14	Glass wool conductivity (λ_{gw})	Slab joist specific heat $(c_{p,sj})$	0.001
15	Slab joist specific heat $(c_{p,sj})$	Thickness of polystyrene Wallmate $(t_{,polW})$	0.0008

Figure 4.16 shows the estimability ranking against the Sobol method ranking of the first 15 parameters. The difference is not very huge. The cluster of the first fifteen parameters is the same except for only one parameter that is the conductivity of glass wool (λ_{gw}), which is ranked 18th by the orthogonalisation method.

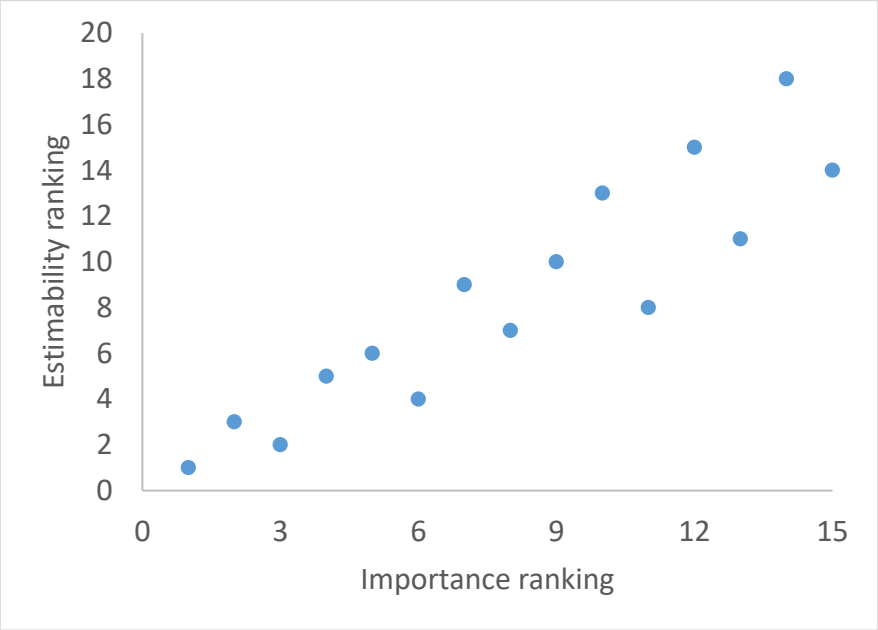

Figure 4.16: Sobol ranking vs orthogonalisation ranking for the first 15 parameters

Calibration with APMC is applied to different sets of parameters starting from the most estimable to the least estimable. Figure 4.17 shows the total identifiability of the parameters for each parameter set: the average over the JS distances over all the parameters is taken. It clearly depicts that the total identifiability decreases with increasing number of parameters as expected. The reason is that the less estimable parameters have lower values of JS. Thus, as more low JS values are included (when including more parameters), the average JS value decreases.

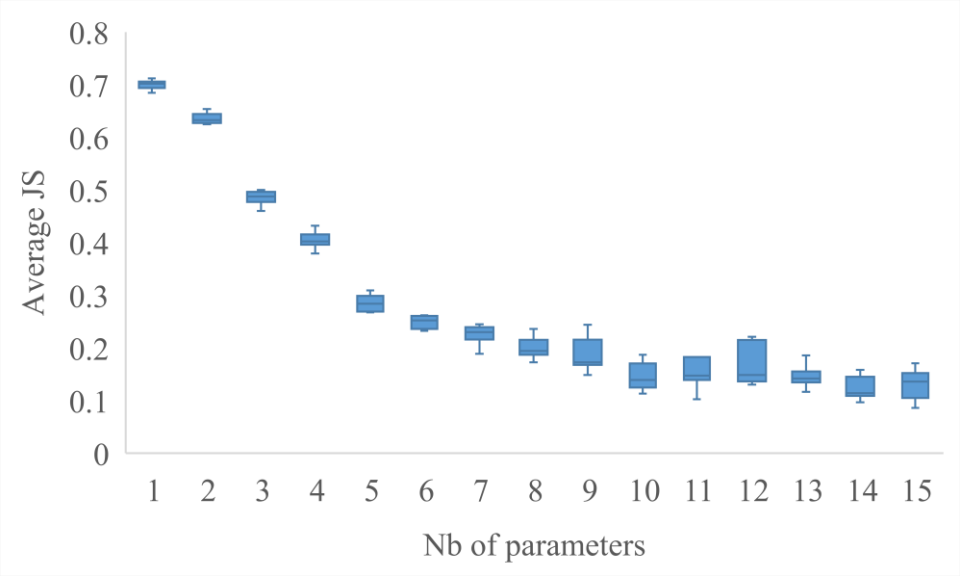

Figure 4.17: Parameters identifiability with increasing parameters number (orthogonalisation ranking)

The decrease in the total identifiability explained by the average JS distance does not necessarily mean that the identifiability of the most estimable parameters decrease. It might mean that the included parameter is very much un-identifiable due to its unimportance and thus it dragged down the total identifiability of the parameters set.

It is thus important to study the identifiability of the most estimable parameter with increasing parameters set size. Figure 4.18 shows the identifiability of the ventilation flowrate. It is clearly depicted that it is mostly identifiable when calibrated alone, however, it does not decrease continuously with increasing number of parameters. Its identifiability depends on the type of parameter that is included. Two significant decreases can be visualised. The first is when the third most estimable parameter (heating power) is included which is expected, since the heat loss through ventilation and the heat gain through the electrical power are included in the same way within the building energy model. In the fifth scenario of the virtual data, only the ventilation flow rate is considered and there is no heating power. This explains why the interaction between these two parameters is not perfectly linear. The second significant decrease is when the fourth parameter (solar albedo) is included. The reason is that in the fifth scenario, the shutters are opened which means that the solar gains will have influence on the interior temperature. Therefore, the solar albedo becomes important and it interacts significantly with the ventilation flowrate.

On the other side, no significant decrease is depicted when the second most estimable parameter (the specific heat of concrete wall) is included. The explanation to this is that the specific heat of the concrete wall is found to be very influential in the second and third scenarios, where the building is in free evolution and there is neither heating power nor ventilation. Consequently, the interaction between the specific heat of concrete wall and the ventilation is not very significant. Similarly, including the sixth, seventh, and eighth most estimable parameter (internal gains, conductivity of the insulation in the ground and the heat transfer coefficient of the window) did not affect the identifiability of the ventilation flow rate for the same reason. However, it is depicted that with an increasing number of parameters, the variability of the JS distance of the ventilation flowrate with different calibration runs increases as clearly shown from the wider interquartile and whiskers.

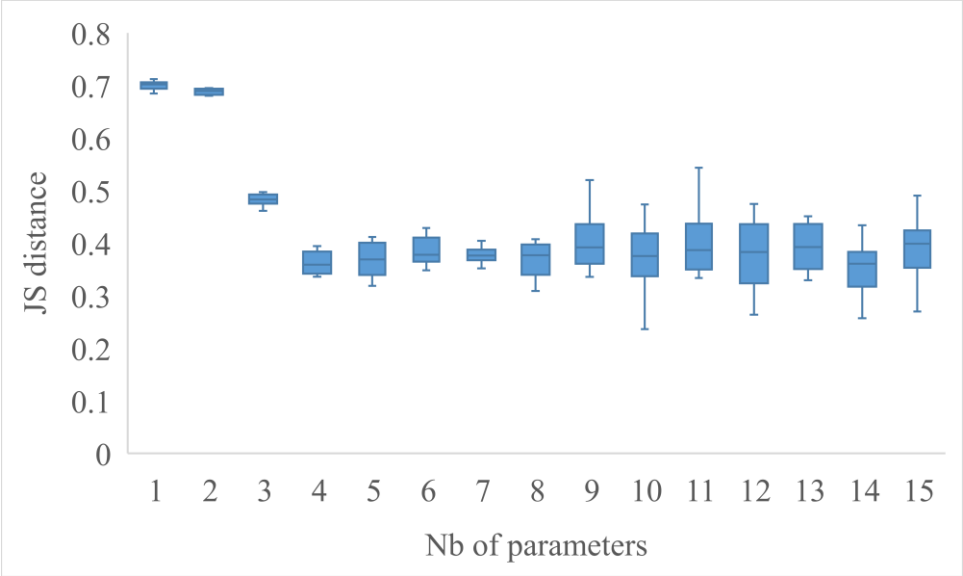

Figure 4.18: Identifiability of the ventilation flowrate

The threshold value used in the orthogonalisation method is analysed from two points of views. The first is the identifiability of the parameters ranked above and below this value, and the second is the model predictive performance after calibrating the parameters ranked above and below this value.

There are six parameters ranked with a threshold of 0.04: ventilation flowrate, heating power, internal gains, solar albedo, the specific heat of concrete, and conductivity of polystyrene wallmate. Figure 4.19 shows the identifiability of each parameter after calibrating a subset of 15 parameters. The ventilation flowrate, heating power, internal gains, solar albedo, and the specific heat of concrete wall are more identifiable than the rest. The conductivity of polystyrene wallmate has slightly better identifiability than the rest. The other parameters are found to be less identifiable. In this figure, there exists no clear clusters of parameters as shown in the case study of the previous section where the parameters that are ranked as the most estimable with a threshold of 0.04 are clearly more identifiable than the rest. This is not the case in this example. However, beyond six parameters, all the other parameters are found to have zero identifiability for some calibration runs: their posteriors are wider than their priors which is explained by retaining a zero for the JS distance as depicted in Figure 4.19. This shows that the parameters ranked above a threshold of 0.04 tend to be more identifiable than the rest.

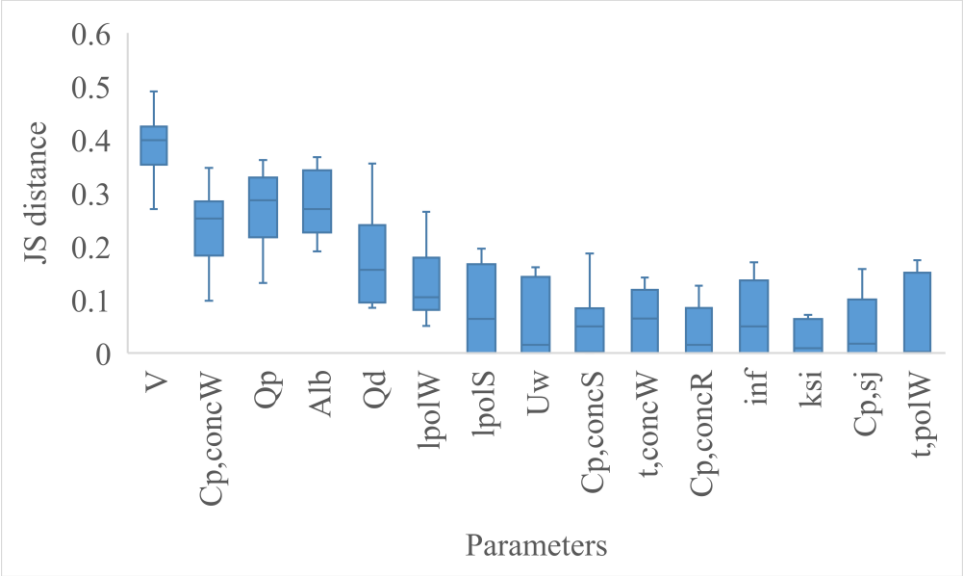

Figure 4.19: Identifiability of parameter subset of 15

Calibrating only the most estimable parameter is the most identifiable calibration, however, other parameters are discarded which affects the capability of the calibrated model to fit well to data. Thus, accounting only for identifiability is not sufficient to have a better model. Even if the identifiability slightly decreases, the addition of another parameter could lead to a better model predictive performance. Figure 4.20 shows the DIC variation as a function of the number of parameters. The propagation of the priors shows the highest values of DIC since no parameter is calibrated yet. A significant decrease in the value of DIC is observed after calibrating the most estimable parameter: it is the largest variation shown. The reason is that the ventilation flowrate (the most estimable parameter) has the most influence on the data. It is observed that the DIC keeps on decreasing until the fourth parameter is included: calibrating only the most estimable parameter is less accurate than calibrating the three most estimable parameters. The DIC of four parameters is slightly higher than that of three parameters. This is explained in the previous paragraph by the decrease in the ventilation flowrate identifiability. Beyond four parameters, the variation in the DIC is not very significant and they are very close to the performance of calibrating three parameters. However, with more than nine parameters, the minimum values of DIC are very low compared to what is depicted with less parameters. Moreover, the maximum values depicted with these sets are relatively high. The reason behind this relatively significant variability as also shown in the identifiability of the ventilation flowrate in Figure 4.18 is that with more parameters, the probability of having different valid combinations of values is higher in addition to the randomness behaviour in the calibration process.

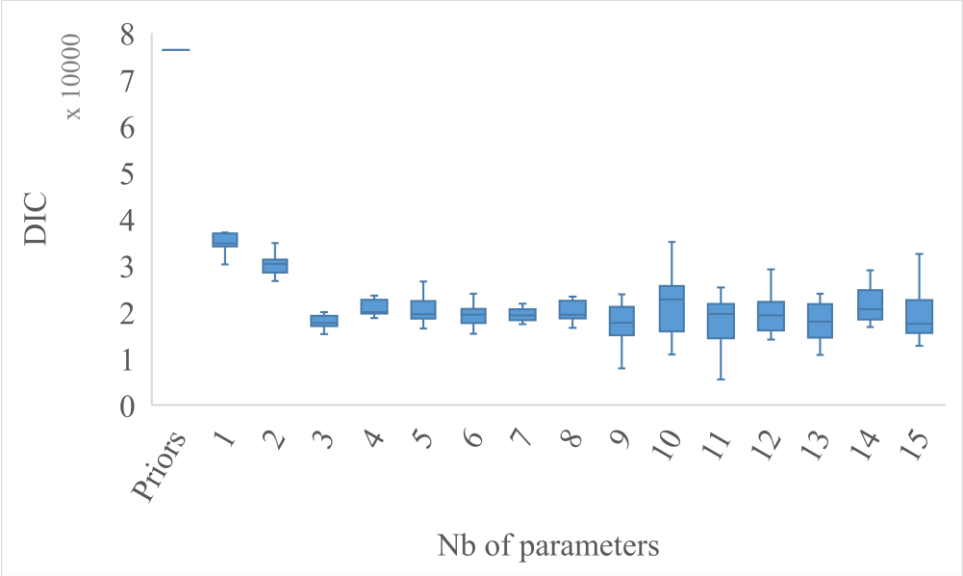

Figure 4.20: Model predictive performance with an increasing number of parameters

It is good to complement the DIC with the RMSE criterion since it gives an additional insight because it represents the difference between the simulations and data (in ℃). In this way, it is easier to draw conclusions about how good or bad the fit is. Figure 4.21 shows the model predictive performance of the calibrated models with an increasing number of parameters in term of RMSE. The same behaviour is depicted as with DIC. For instance, for the subset of 10 parameters, the lower whisker corresponds to RMSE of 0.102℃, while the upper whisker corresponds to an RMSE of 0.15℃. This difference is not negligible; especially, the RMSE of the priors propagation is 0.2℃. To go further, the calibration runs corresponding to these two whiskers are investigated more in depth: their posteriors estimation and the predicted temperature profile are analysed in the following paragraph. These two runs are retained since they seem to record the highest difference in RMSE and DIC among all other sets.

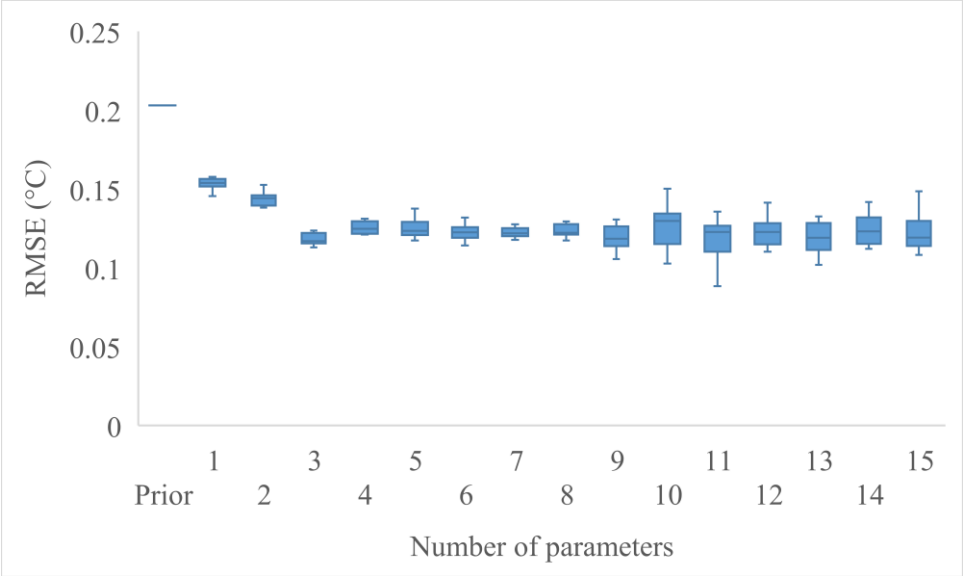

Figure 4.21: Model predictive performance in RMSE

Figure 4.22 shows the posteriors for these two calibration runs, the posteriors on the left corresponds to the maximum DIC, and the ones on the right correspond to the minimum DIC. It is clearly depicted that the posteriors on the left are more uncertain and some parameters have slightly wider posteriors than the priors, which means that samples that are further from the true values are retained in the posteriors. On the contrary, the posteriors on the right are narrower towards the true values, which means that the retained samples are mostly closer the true values. Given that the convergence criterion for all the calibration runs is similar, this means that both posteriors yielded similar accuracy during calibration. This highlights the presence of overparametrisation problem occurring with these sets (subsets of more than eight parameters). Overparametrisation is more significant in the figures on the left than in the figures on the right.

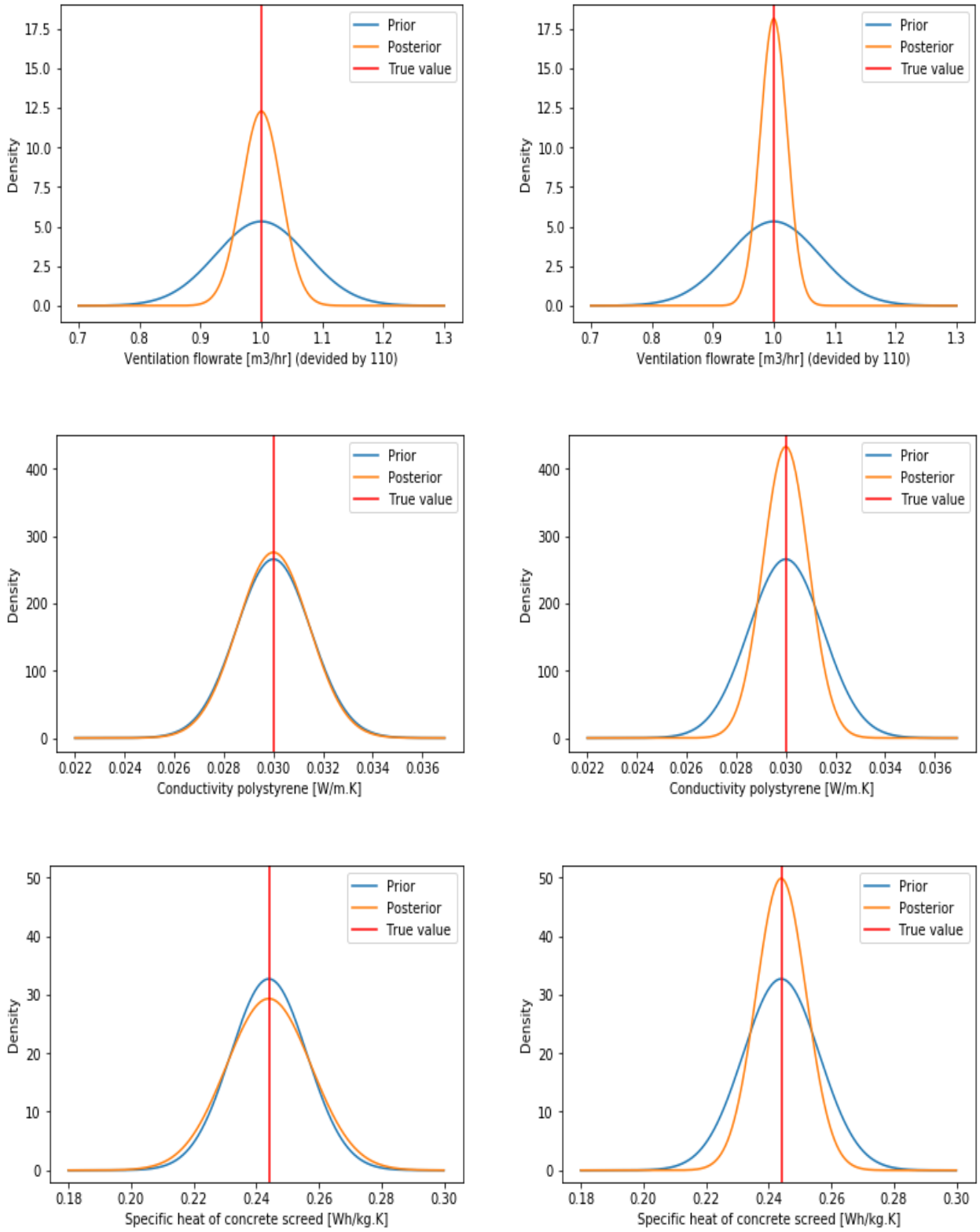

Figure 4.22: Posteriors vs priors in two calibration runs with a subset of ten parameters (left: high DIC; right: low DIC)

Having visualised the variabilities in larger sets of parameters in identifying the posteriors, it is important to assess how significant this variability is in estimating the quantity of interest,

which is in this case the temperature profile. The RMSE values showed that these variabilities could be significant, but it is better to depict the temperature profile. The posteriors belonging to the two calibration runs presented in Figure 4.22 are propagated (Figure 4.23). The difference between the two calibration runs appears not to be substantial: the two propagated curves belonging to calibration runs 1 and 2 are close to each other. This is actually due to the fact that the priors themselves are close to the virtual data (RMSE 0.2℃). In fact, the two propagated curves are significantly different if you look at them relative to the prior propagation (especially for the first two scenarios).

Accounting only for the posteriors propagation for the two runs without looking at the prior, one can say that the difference in the predicted temperature profile is not significant. However, compared to the prior, the reduction in the RMSE is non negligible. Accordingly, it could be said that the variability present in the set of ten parameters is not negligible.

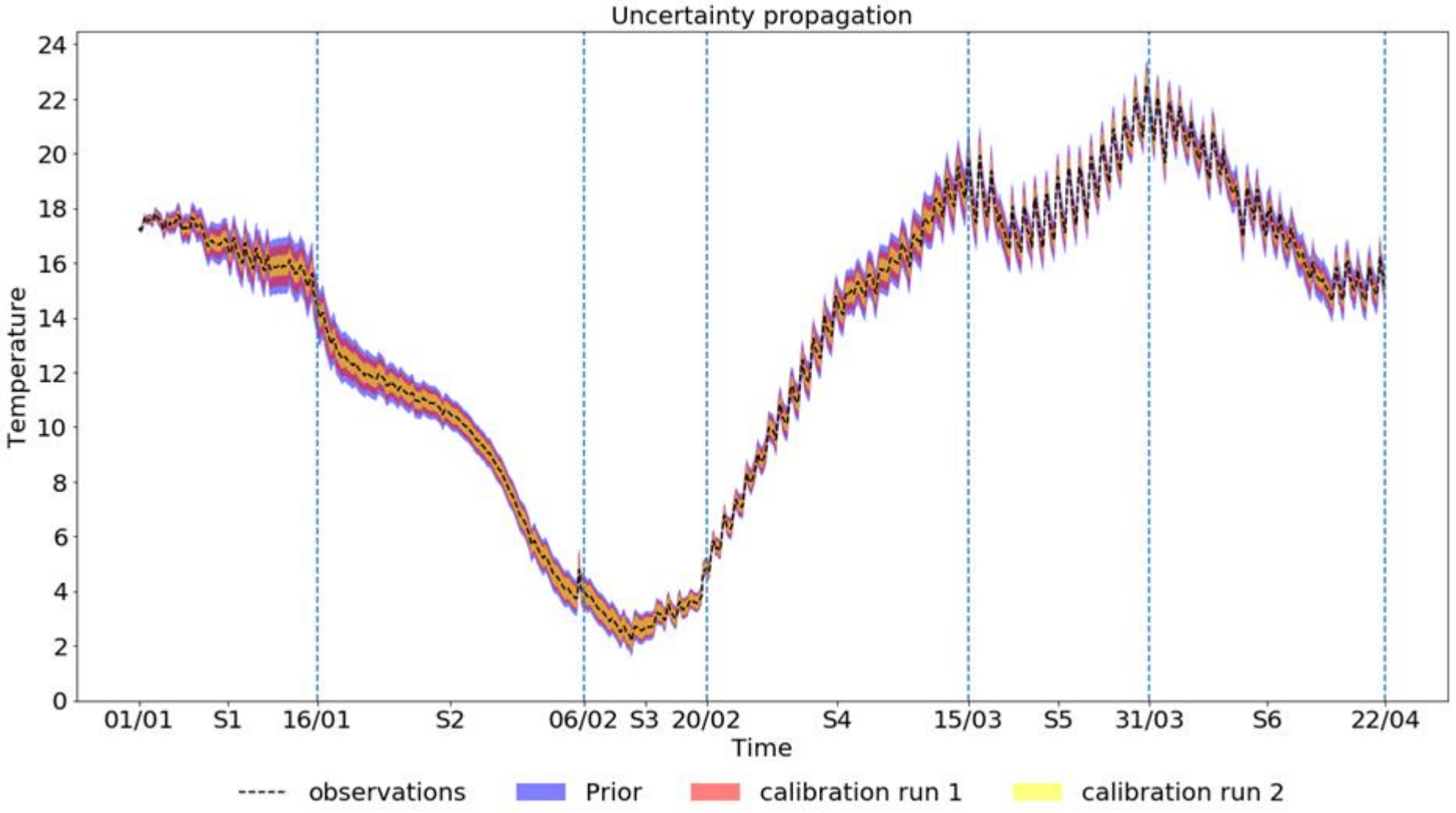

Figure 4.23: Propagation of prior and posterior of two calibration runs in set of 10 parameters

To summarise, calibrating up to eight parameters in this case study showed no significant variabilities and randomness. In terms of model predictive performance, the three most estimable parameters (ventilation flowrate, specific heat of concrete wall and the heating power) that are ranked with a threshold of 0.427 showed the best performance, even though it is not significantly better than calibrating more parameters. For example the average value in the box plot of the set of three parameters lies within the interquartile of the box plot of the six parameters, and the minimum value of both is similar. There is no significant difference between calibrating 4, 5, 6, 7, and 8 parameters: the corresponding DICs and RMSEs are very close.

Thus, it could be said that for a stable calibration performance with repetition, calibrating up to eight parameters in this case study is reasonable where the randomness effect of calibration is not significant, the best performance is with three parameters. The threshold value (0.04) showed no clear relation to the model predictive performance. Stability and similar performance is observed when calibrating parameters ranked below and above 0.04. More case studies need to be analysed in order to either set a general recommendation for the threshold value or to show that it varies with different cases.

4.3.3 Conclusion

In this section, the effect of the number of parameters on the calibration results in controlled conditions is analysed. The parameters are ranked based on the results of the sensitivity-based identifiability analysis. Calibration is executed sequentially, firstly, the most estimable parameter is calibrated, then the two most estimable parameters are calibrated and so on until fifteen parameters are included.

The effect of the parameters number is evaluated based on the identifiability in the posteriors using the JS distance on the one hand and on the calibrated model predictive performance using the DIC criterion on the other hand. The first six estimable parameters are shown to be more identifiable than the rest, however, the three most estimable parameters yielded the best model predictive performance. Beyond eight parameters, the calibration results started to be more variable due to the more possible parameters combinations. Accordingly, for this case study, three parameters could be considered as a good choice for the number of parameters. Eight parameters are also a good choice since the difference between their corresponding results and the results of calibrating three parameters is not very significant. A

threshold value of 0.04 is convenient for this case study to distinguish the identifiable parameters from less identifiable ones, but in terms of model predictive performance, calibrating parameters ranked below this threshold showed similar performance. Thus, further analysis should be conducted on this value.

4.4 Chapter conclusion

The objective in the first part of the chapter is to analyse the importance of ranking the parameters in terms of estimability compared to ranking them in terms of importance, and how each influences the calibration results. For this, Sobol's method is used for importance ranking, and then orthogonalisation method is considered for estimability ranking. The study showed that calibrating the most estimable parameters yielded better results in terms of identifiability and in terms of fitting to data than calibrating the most influential parameters.

It should be noted that in this study, when the most estimable parameters are taken into account, the influential parameters that are discarded are set at their true values. In reality, the discarded parameters could be uncertain and they could be set at false values. Accordingly, this might affect the ability of the calibrated model to fit to the testing data if its conditions are different than the training data. The current study did not tackle this issue because perfectly controlled conditions were assumed. A possible approach to study this would be to fix the discarded parameters at false values and repeat the same methodology.

In the second part of the chapter, the objective is to analyse the performance of calibration with different numbers of parameters. In the retained case study, three parameters yielded the best performance. Calibrating eight parameters is not significantly worse than calibrating only three parameters. More parameters showed significant variabilities in the calibration results with repetitions. It is also concluded that there is no clear relations between the orthogonalisation method threshold and the model predictive performance of calibration, however, a value of 0.04 is found to provide a reasonable performance: it does not include too many parameters, which would yield worse performance.

Chapter 5

Adaptive random forest

In chapter 3, it was shown that ABC-RF is able to yield better performance than other methods with small number of model evaluations. In this chapter more investigation is done on this method. Accordingly, a new approach called adaptive random forest (ARF) that is based on ABC-RF is proposed. Detailed explanation about the principle of ARF is firstly provided. Then it is applied on different virtual case studies and compared with the methods used in chapter 3. Finally, all the methods used in this thesis are applied on a real case study with on-site measurements.

Résumé du chapitre

Les méthodes bayésiennes existantes nécessitent un nombre non négligeable de calculs de vraisemblance, c'est-à-dire un nombre non négligeable d'évaluations de modèles. Cela les rend inefficaces sur le plan informatique, sauf si un métamodèle est constitué pour remplacer l'original, au prix d'une erreur et d'une incertitude supplémentaires, ce qui affecte la précision de l'ensemble du processus de calibrage. Ainsi, il est intéressant de converger vers des résultats précis avec un nombre réduit de calculs de vraisemblance. ABC-RF, une méthode récemment introduite, semble selon les résultats du chapitre 3 avoir un potentiel d'approximation des distributions a posteriori intéressant, avec un nombre limité d'évaluations de modèles. Cependant, elle devient moins précise que les autres méthodes avec davantage de simulations.

Dans ce chapitre, l'objectif est d'étudier des solutions aux problèmes des méthodes de calibrage, en particulier le problème lié à la charge de calcul. A cette fin, en se basant sur les développements récents dans le domaine et particulièrement sur l'association entre le calibrage bayésien et les forêts aléatoires (ABC-RF), un nouvel algorithme appelé forêt aléatoire adaptative (ARF) est proposé. Cet algorithme bénéficie de ABC-RF car il ne nécessite pas la définition de nombreux hyper-paramètres, et de l'échantillonnage séquentiel de Monte Carlo car il échantillonne à partir de distributions plus proches de la postérieure plutôt que d'échantillonner à partir des distributions initiales (priors), ce qui le fait converger vers les postérieurs avec moins d'évaluations de modèles. De plus, ARF, avec l'échantillonnage séquentiel adapté, évite la limitation d'extrapolation d'ABC-RF.

La méthode proposée est appliquée sur une étude de cas virtuelle avec cinq distributions initiales différents : des distributions larges, des distributions précises et des distributions décalées des vraies valeurs. Les indicateurs utilisés sont les mêmes que ceux utilisés au chapitre 3 lors de la comparaison de différentes méthodes bayésiennes. Les résultats ont montré que cette méthode peut obtenir des estimations très précises des vraies valeurs et par conséquent une bonne performance prédictive du modèle avec un petit nombre d'évaluations du modèle (pas plus de 3000). La méthode est également comparée à celles présentées au chapitre 3 et elle a montré une efficacité de calcul considérablement meilleure (plus de 10 fois plus rapide que certaines autres méthodes).

Enfin, toutes les méthodes sont appliquées à un cas d'étude réel avec un profil de température mesuré expérimentalement. Le calibrage a permis d'améliorer la performance

prédictive du modèle sur les données d'entraînement et de test avec tous les algorithmes utilisés dans ce chapitre. Des performances relatives similaires se retrouvent donc aussi bien sur des données réelles que sur des données virtuelles. Cependant, ARF s'avère diverger avec un nombre croissant d'itérations, ce qui n'est pas le cas des autres algorithmes. Mais dans cette évaluation, le nombre d'arbres et la taille des feuilles n'ont pas été augmentés en fonction du nombre de simulations. Davantage d'investigations doivent être effectuées sur cette méthode d'ARF pour améliorer ses performances dans des conditions non contrôlées, et vérifier sa convergence. Pour le moment, en l'absence de cette vérification, APMC pourrait être le meilleur choix parmi ceux illustrés dans cette thèse. ARF a montré un potentiel par rapport aux autres méthodes avec un nombre limité d'évaluations, mais cela doit être confirmé à partir d'autres études de cas.

5.1 Introduction

Existing Bayesian methods require a non-negligible number of likelihood computations, which means a non-negligible number of model evaluations. This makes them computationally inefficient except if a metamodel is trained to replace the original one, under the cost of adding additional error and uncertainty which affects the accuracy of the whole calibration process. Thus, it is interesting if precision could be reached with a reduced number of likelihood computation. ABC-RF, a recently introduced method, is shown in chapter 3 having a potential in approximating the posteriors with a limited number of model evaluations. Moreover, ABC-RF overcomes the difficulty of choosing appropriate summary statistics for the data, and it does not require the definition of a distance function or a minimum threshold as it is the case for other ABC methods. However, with ABC-RF all samples are generated from the priors and cannot be extrapolated outside the ranges of the priors.

In this chapter, the objective is to tackle the disadvantages inherited with the calibration methods particularly the problem related to computational burden. To this end, based on recent development in the field and particularly on ABC-RF, a new algorithm called adaptive random forest (ARF) is proposed. This algorithm benefits from ABC-RF because it does not require the definition of many hyper-parameters, and from sequential Monte Carlo sampling because it samples from distributions closer to the posterior rather than sampling from the priors, which makes it converge to the posteriors with less model evaluations. Moreover, ARF, with the adapted sequential sampling, avoids the extrapolation limitation of ABC-RF.

The proposed method is applied to virtual and real in-situ data and its performance is compared to other calibration methods presented in chapter 3 in terms of accuracy and computational efficiency.

5.2 Motivation

ABC-RF has a significant advantage against other algorithms: it requires less hyper-parameters and these parameters are easier to tune. The hyper-parameters in ABC-RF are related to those required to train the random forest such as the number of trees, leaf size, etc. In fact, there exists numerous applications of random forests in the literature, including some clear recommendations concerning these hyper-parameter. Moreover, these hyper-parameters can easily be tuned. One can train as many random forests as needed for tuning purposes without

the need to launch any additional BEM simulation. However, the structure of the other approaches necessitates to sample and go back to the model iteratively which requires a whole new set of model evaluations. One may say that this could be avoided if the original model was replaced by a metamodel and thus, tuning would become possible with only one data set generated just like ABC-RF. However, it can be argued that using a metamodel makes the whole problem more uncertain and using the original model is definitely favoured. ABC-RF also avoids the difficulty inherited in ABC approaches to select a small number of sufficient summary statistics. This is considered one of the main advantages of ABC-RF over other algorithms. This is why it is mentioned in this section, however, the selection of sufficient summary statistics is not a very critical problem with BEM applications in the case of time series data, where the RMSE could be a sufficient.

There are some problems with ABC-RF concerning the accuracy of the generated posteriors and their variability with repetitions. Despite this, it is shown in chapter 3 that ABC-RF has a potential with relatively small data set size (Akkari et al. 2022). In this section, a new method, adaptive random forest (ARF), based on ABC-RF is proposed. This method applies the sequential sampling techniques to ABC-RF so that more samples from the posterior regions are generated. Accordingly, ARF benefits from all the advantages of ABC-RF and solves its variability and accuracy issues, but under the cost of adding some new hyper-parameters that are detailed in the following sections.

5.3 Principle

Random forests require large data sets to solve the variability issues that it might suffer from. Raynal et al. (2017) recommended a default choice of 100,000 data set samples and suggested to consider a larger data set if the variabilities are non-negligible. This might decrease the variabilities but under the cost of computational burden. A better way is that, instead of generating the data set in one batch, it could be generated adaptively. The algorithm is initialised by training a random forest on a data set $\{\theta_{1:P}^1, \dots, \theta_{1:P}^{N^1}; y^1, \dots, y^{N^1}\}$ of size N^1. The expectation and variance of the parameters distributions are extracted like in ABC-RF. This serves as the first iteration of the algorithm. In the subsequent iterations $t = \{2, \dots, T\}$, new samples $\theta^t = (\theta_{1:P}^1, \dots, \theta_{1:P}^{n^t})$ of size n^t are generated from the distributions of the previous iteration $(t-1)$. These samples θ^t with their corresponding outputs $y^t = (y^1, \dots, y^{n^t})$ are then concatenated

with the samples of all the preceding iterations to form a new data set of size N^t. This allows to adaptively generate samples from regions closer to the posteriors than to the priors.

In fact, ARF does not generate samples directly from the parameters distributions: the samples are drawn from uniform PDFs constructed over the $\pm 3\ \sigma$ of these distributions ($\pm 3\ \sigma$ corresponds to 99.7 % confidence interval). The underlying reason is that sampling from the tails of a normal distribution is significantly less probable. Instead, if it were replaced with a uniform distribution constructed over its $\pm 3\ \sigma$, it would become more probable to draw the samples that are in the tails. This is important if the true value of the parameter lies in the tails of the priors. One can say that if a sufficient number of samples are generated, the tails could be explored. This could be argued that the aim of ARF is to reach convergence with as few model evaluations as possible, and this is why with ARF a small number of samples are generated at each iteration, so a uniform distribution is a better choice to well explore the whole parameter space even in the tails. In this thesis, the boundaries of the uniform distribution is called the sampling bounds, and the value that is multiplied by the standard deviation is called the search parameter.

Choosing a fixed value for the search parameter makes the algorithm prone to the problem of extrapolation: it might be difficult to draw samples outside the range of the priors. This could be problematic if the true value of the parameter is not inside the prior range. To overcome this issue, the sampling bounds of the uniform distribution from which samples are drawn at each iteration are adaptively modified as follows:

$$U^t[B_l^t, B_r^t] = \begin{cases} \left[\mu^t(Y) - k^t \cdot \sqrt{V^t(Y)}\,;\ \mu^t(Y) + 3 \cdot \sqrt{V^t(Y)}\right] & if\ x^t < 0.5 \\ \left[\mu^t(Y) - 3 \cdot \sqrt{V^t(Y)}\,;\ \mu^t(Y) + k^t \cdot \sqrt{V^t(Y)}\right] & if\ x^t > 0.5 \end{cases} \quad (5.1)$$

where B_l and B_r are the left and right sampling bounds of the uniform distribution $U^t[B_l, B_r]$. x^t represents the distance from the expected distribution value estimated at the current iteration $\mu^t(Y)$ to the sampling bounds of the distribution of the previous iteration $[B_l^{t-1}, B_r^{t-1}]$. The distance is then normalised by the difference between the sampling bounds as follows:

$$x^t = -\frac{\mu^t(Y) - B_l^{t-1}}{B_r^{t-1} - B_l^{t-1}} \quad (5.2)$$

k is the search parameter. Its default value is 3 since it accounts for 99.7 % confidence interval as introduced previously. The search parameter is used to identify the sampling bounds of the

current iteration: the bounds between which the samples of the subsequent iteration $t+1$ are generated. Depending on the value of x^t, the search parameter is identified. For example, if $\mu^t(Y)$ is closer to the left bound B_l^{t-1} which is represented by $x^t < 0.5$ in equation (5.1), the default value of the search parameter is retained to determine the right bound B_r^t, and the left bound B_l^t is determined by setting a non-default value for k. In this case, the value of k is computed from the following expression:

$$k^t = -\frac{A_1 - A_2}{2} \times \sin(\arctan(ax - c)) + \frac{A_1 + A_2}{2} \tag{5.3}$$

This expression is used since it allows for *k* to change between two bounds A_1 and A_2 following a monotonic nonlinear decrease with increasing *x*, which allows to widen or shorten the sampling bounds according to how close or far the distribution expected value is to the boundaries. A_1 and A_2 are the lower and upper asymptotes. The reason behind using a formula that has an upper and a lower bounds is that k should neither be too small, which would induce the risk that the parameters space is not well explored nor to be too large with the risk that more samples and iterations are required which affects the computational efficiency of the algorithm. a and c are the slope and location parameters. These parameters identify how the search parameter k changes with x.

The incorporation of an adaptively identified search parameter is essential to widen the sampling bounds and narrow them when necessary, so that the algorithm does not get stuck in local minimums. It also ensures that ARF can extrapolate outside the ranges of the priors. If the variances $V^t(Y)$ and k^t at a current iteration are both large, this will result in a very wide sampling bounds, which could be computationally problematic. Accordingly, the upper asymptote A_1 is also updated adaptively at each iteration. The idea is that if the variance is very close to the variance of the prior, A_1 could be reduced to $\hat{A}_1$ as follows:

$$\hat{A}_1^t = A_1\left(1 - \frac{\sqrt{V^t(Y)}}{\sqrt{V^{prior}}}\right) \tag{5.4}$$

If the distribution at iteration t is very similar to that of the prior such that $\hat{A}_1^t$ was found to be smaller than the lower asymptote A_2, then A_2 is chosen as the value for $\hat{A}_1^t$. Moreover, if k^t is found bigger than $\hat{A}_1^t$, then $\hat{A}_1^t$will be retained as the value for k^t. This modification ensures that the sampling bounds are not very wide while at the same time aiming at sufficient parameter exploration.

The steps of the proposed method are listed in algorithm 5.1. ARF is initialised with the prior distributions without the application of equations (5.3), and (5.4). Starting from the third iteration, these equations are introduced. The reason is that ARF without these improvements is capable of finding the true values of the parameter even if they lie in the boundaries of the priors. ARF is able to push the boundaries a little and find those values without the need to introduce the search parameter. In this case, introducing the search parameter increases the parameter space from which to sample which causes a higher computational cost. Thus, to study the possibility that the true value is in the tails of the priors, the algorithm is given time to sample from the tails in the second iteration without any tuning of the boundaries, and then the 3rd and all the subsequent iterations are generated after introducing the search parameter.

Algorithm 5.1

1. Initialise ARF with N^1, sequence for n^t, and k parameters.
2. Generate the data set $\{\theta_{1:P,1}, \dots, \theta_{1:P,N^1};\ y_1, \dots, y_{N^1}\}$ and train RF.
3. Compute $\mu^1(Y)$ and $V^1(Y)$.
4. Sample $\{\theta_{1:P,1}, \dots, \theta_{1:P,n^2}; y_1, \dots, y_{n^2}\}$ from $U\left[\mu^1(Y) - 3 \cdot \sqrt{V^1(Y)}; \mu^1(Y) + 3 \cdot \sqrt{V^1(Y)}\right]$.
5. **for** t in $\{2, \dots, T\}$:
   a. Concatenate data of all iterations $\{\theta_{1:P,1}, \dots, \theta_{1:P,N^t};\ y_1, \dots, y_{N^t}\}$.
   b. Train RF and compute $\mu^t(Y)$ and $V^t(Y)$.
   c. Compute k^t and $\hat{A}_1^t$.
   d. Sample $\{\theta_{1:P,1}, \dots, \theta_{1:P,n^t}; y_1, \dots, y_{n^t}\}$ from $U^t[B_l, B_r]$.
   e. Increment $t = t + 1$.

5.4 Parameters tuning

The hyper-parameters related to the construction of the random forest are not tuned. Different studies have been performed to analyse the effect of the random forest hyper-parameters (Genuer et al. 2008; Genuer et al. 2010; Biau and Scornet 2015). Here, the values for the hyper-parameters recommended by Raynal et al. (2017) that comply with the general recommendations in RF context are retained. Table 5.1 shows the values considered for the random forest hyper-parameters.

Table 5.1: ABC-RF hyper-parameters

Name	Symbol	Value
Number of trees	b	500
Maximum leaf size	N_{min}	5
Number of features at each split	n_{try}	nb_features/3

The data set size is another hyper-parameter for the random forest, however, in the context of ARF, it is replaced with another hyper-parameter related to sample size of the initialising data set N^1: the number of samples generated to initialise the algorithm at the first iteration. To ensure a good exploration of the parameter space, it is set at 1000. The sample size at all the subsequent iterations in ARF is defined by n^t. To explore well the space and at the same time to reduce the computational cost of the method, 300 samples are generated at each iteration. Some preliminary analysis on this parameter not presented here showed that less samples at each iteration is adequate to reach same convergence with a smaller number of simulations.

There are additional tuning parameters specific to the application of ARF (Table 5.2). The default value of the search parameter is set at 3, since $\pm 3\sigma$ corresponds to 99.7 % confidence interval which is sufficient to explore the distribution. The default values are used for the boundary to which the distribution is further away and an adaptively computed value of the search parameter is used to the other boundary. For instance, if the distribution moved to the right, the default values are used on the left boundary and the adaptively computed values are used on the right boundary. For this adaptive computation, the hyper-parameters: $A_1, A_2, a,$ and c are introduced. A_1 should neither be too large which would result in a very wide sampling bounds, nor too small which would make the search parameter very close to its default value (3), and thus the benefit of the search parameter would be lost. Accordingly, a value of 9 is found to be an adequate choice. A_2 is set at the default value of the search parameter that is 3 to ensure a good exploration of the distribution at all stages and iterations of the algorithm, since $\pm 3\sigma$ corresponds to the 99.7% confidence interval as mentioned earlier. $a,$ and c are related to the shape of the search parameter function. Their values (29 and 4 respectively) are chosen in a way that the k is nearly constant at a value close to A_2 when the estimated distribution is not very close to the sampling bound and quickly increases otherwise. Figure 5.1 depicts the evolution of the search parameter k against the distance (represented by x in the figure) from the distribution expected value to the closer bound. The function shape depicted corresponds to the hyper-parameters values discussed and shown in Table 5.2. These values

should be robust against different case studies even in different fields since they are identified based on reasonable justification and not randomly.

Table 5.2: ARF hyper-parameters

Name	Symbol	Value
Sample size (initialisation)	N^1	1000
Sample size (at each iteration)	n^t	300
Upper asymptote of the k function	A_1	9
Lower asymptote of the k function	A_2	3
Slope of the k function	a	29
Shape parameter of the k function	c	4

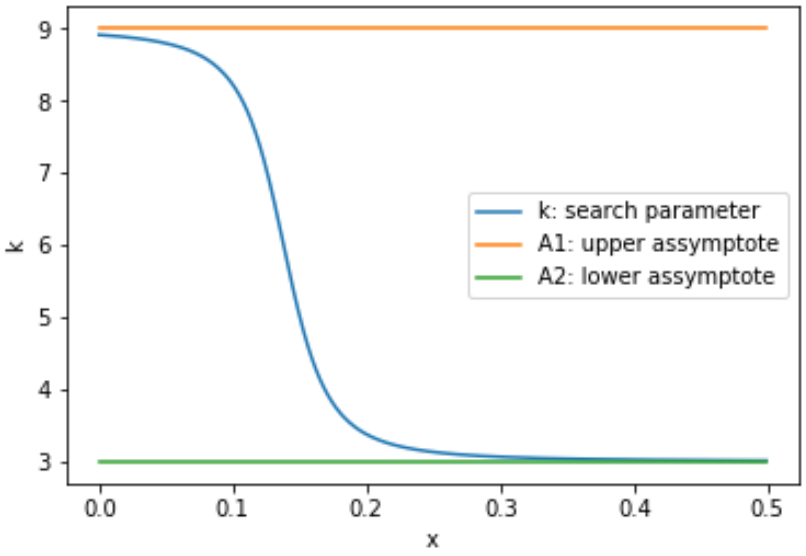

Figure 5.1: Search parameter (k) (illustration of equation-5.3)

5.5 Validation in controlled conditions

The proposed method is applied firstly in controlled conditions. This means that the data on which calibration is performed are generated virtually from the BEM under known parameters values that are called true values. The calibration role is to find these true values using the generated virtual data. Uncertainties regarding the model and measurements are thus avoided. This application is presented in section 5.5. In a second step, section 5.6, ARF is applied to a real case study using on site measurements.

5.5.1 Methodology and criteria

The proposed method is validated on virtual data, and compared with other algorithms. Accordingly, different cases corresponding to different priors' definitions are considered. Some cases are generated with narrow priors and others with wider priors. In all the cases studies, the prior means are shifted from the true values of the parameters as follows:

$$\mu_\theta = \theta_{true} \pm s\sigma_\theta \tag{5.5}$$

where θ_{true} is the true value of the parameter, s is the shift from the true value, and σ_θ is the standard deviation of the prior. In certain cases, the priors are shifted so that the true values are outside the 99.7% confidence region. This allows validating the method performance in interpolating and extrapolating outside the priors' ranges.

The comparison criteria used in chapter 3 – normalised Euclidean distance between the true values and the posteriors (d_{dist}), and the average RMSE between the posteriors propagation and the virtual data – are retained. Likewise, these criteria are not only evaluated on the posterior distributions, but also on the distributions generated at each iteration of the algorithms. This allows to evaluate the performance of each algorithm with an increasing number of model evaluations which allows for a more comprehensive comparison.

5.5.2 Case study

The case study presented in chapter 2 is retained in this study. The sensitivity-based identifiability analysis is applied on this case study and illustrated previously (refer section 4.2). The first six most estimable parameters - ventilation flow rate (Q_v), internal gains (Q_d), heating power (Q_p), specific heat of concrete ($c_{p,c}$), solar albedo (alb), and the conductivity of polystyrene (λ_P). are chosen. Five different cases are considered in this section. The difference between the cases is in the selection of the priors.

In the first two cases as presented in Table 5.3, the priors are shifted from the true values by 1 and 2.9 standard deviations respectively. The priors are not all shifted in the same direction: all to the right or all to the left of the true value. This enables to avoid different possible bias and interactions. In case 3, some of the parameters are shifted in a way that the true values are located outside the ranges of the priors while the others are within the prior ranges. In case 4, all the parameters are shifted by four standard deviations.

Table 5.4 presents the standard deviations of the priors used in cases 3 and 4 which are narrower than those used in the previous cases to check if the algorithm might get stuck in local minimums. It also shows by how many standard deviations the prior of each parameter is shifted from the true value. Case 5 is used to perform a comparative analysis between the algorithm presented in chapter 3 and ARF. For this case, the virtual data are generated based on different values and the priors are shifted by two standard deviations from the true values (Table 5.5).

For all the cases studied, the parameters that are not included in calibration are set at their true values used for generating the virtual data which means that theoretically, in this case, since there is no uncertainties about the data used and no model error, the calibration should converge to the true values.

Table 5.3: Prior distributions (cases 1 and 2)

Parameters	Prior distribution $N(\mu_\theta, \sigma_\theta)$		True value	Unit
	Case 1 ($s = 1$)	Case 2 ($s = 2.9$)		
Q_v	$N(88,22)$	$N(46,22)$	110	$[m^3/h]$
Q_d	$N(166,41.6)$	$N(87,41.6)$	208	$[W]$
Q_p	$N(1080,120)$	$N(852,120)$	1200	$[W]$
c_{pc}	$N(0.362,0.106)$	$N(0.5634,0.106)$	0.256	$[Wh/(Kg.K)$
alb	$N(0.55,0.2)$	$N(0.93,0.2)$	0.35	$[-]$
λ_P	$N(0.05,0.02)$	$N(0.088,0.02)$	0.03	$[W/(m.K)]$

Table 5.4: Prior distributions (cases 3 and 4)

	Prior distributions shift (s) and standard deviation (σ_θ)				
	Case 3		Case 4		
Parameters	s	σ	s	σ	Units
Q_v	2	11	4	11	$[m^3/h]$
Q_d	2	20.8	-4	20.8	$[W]$
Q_p	6.25	20	4	20	$[W]$
c_{pc}	-4.2	0.0256	-4	0.0256	$[Wh/(Kg.K)]$
alb	-5.7	0.035	4	0.035	$[-]$
λ_P	6.6	0.003	-4	0.003	$[W/(m.K)]$

Table 5.5: Prior distributions (case 5)

Parameters	Prior distribution $N(\mu_\theta, \sigma_\theta)$	True value	Unit
	Case 5 ($s = 2$)		
Q_v	$N(110,11)$	132	$[m^3/h]$
Q_d	$N(208,20.8)$	250	$[W]$
Q_p	$N(1200,20)$	1161	$[W]$
c_{pc}	$N(0.256,0.0256)$	0.3072	$[Wh/(Kg.K)$
alb	$N(0.35,0.035)$	0.42	$[-]$
λ_P	$N(0.03,0.003)$	0.024	$[W/(m.K)]$

5.5.3 Application and results

In all the cases, ARF is terminated at the 7th iteration since a sufficient accuracy is established at this level, which corresponds to a total of 2800 model evaluations. Sufficient accuracy here is identified from the relative decrease in RMSE compared to the priors, and how close the posteriors are to the true values. Please note that this sufficiency is identified only visually without considering quantitative accuracy thresholds for the RMSE or the posterior distributions.

Figure 5.2 and Figure 5.3 show the evolution in the estimation of the heating power with the algorithms iterations for cases 1 and 2 respectively for ARF. Only one parameter is depicted (heating power) to illustrate the performance of ARF on the parameters estimation. The other parameters are all well estimated and illustrating them gives no additional information regarding the performance of ARF. The reader can refer to appendix E to visualise the estimation of all other parameters in the various cases. The green line in Figure 5.2 and Figure 5.3 is the expected value at each iteration and the other solid lines are the corresponding lower and upper bounds. The dotted line is the represents the true value of the parameter. The y-axis is the values of the parameter but given in terms of ratio: 1 represents the true value which is 1200 W for heating power. The posteriors in these two figures are centred on the true value. The standard deviation in case 1 and case 2 are 0.0067 and 0.009 respectively compared to 0.1 for the prior.

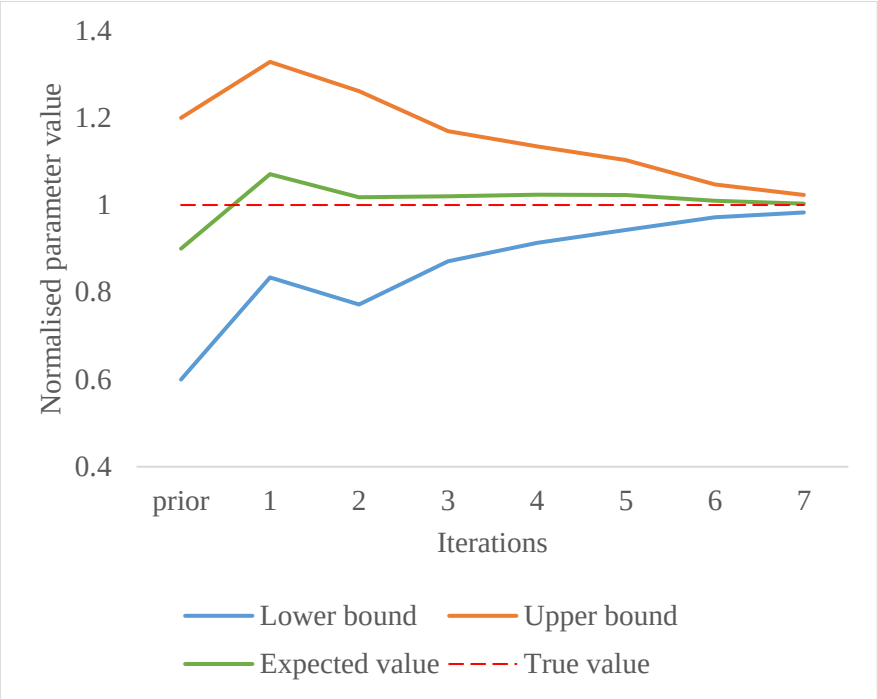

Figure 5.2: Evolution of heating power with ARF iterations (case 1)

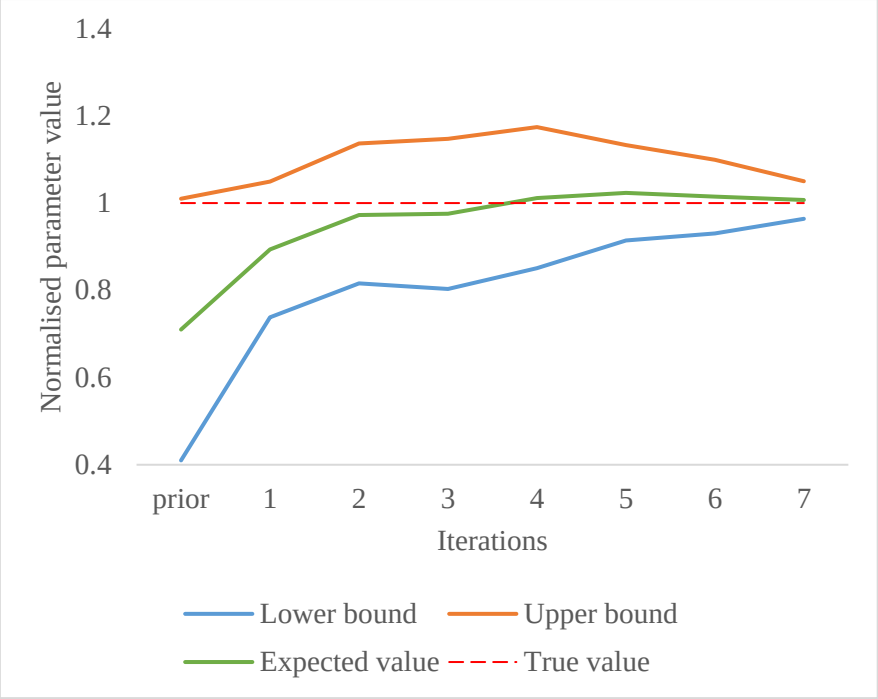

Figure 5.3: Evolution of heating power with ARF iterations (Case 2)

ARF is able to get close to the true value after the first iteration, however with a relatively large variance. With the subsequent iterations, ARF is able to precisely identify the true value with smaller variance. These results were attained for both cases with only 7 iterations corresponding to 2800 model evaluations. Figure 5.4 shows the variation of the search parameter with respect to the algorithm iterations. It shows that the value of the search

parameter is almost constant at the default value for all the iterations. Even if the search parameter is not adaptively tuned and is kept at the default value 3, it is able to well estimate the parameters even if the true values are at the boundaries of the priors. This justifies the benefit of using the default value of the search parameter for the first two iterations, where ARF is capable of independently managing the boundaries if the true value is not very far from the priors range.

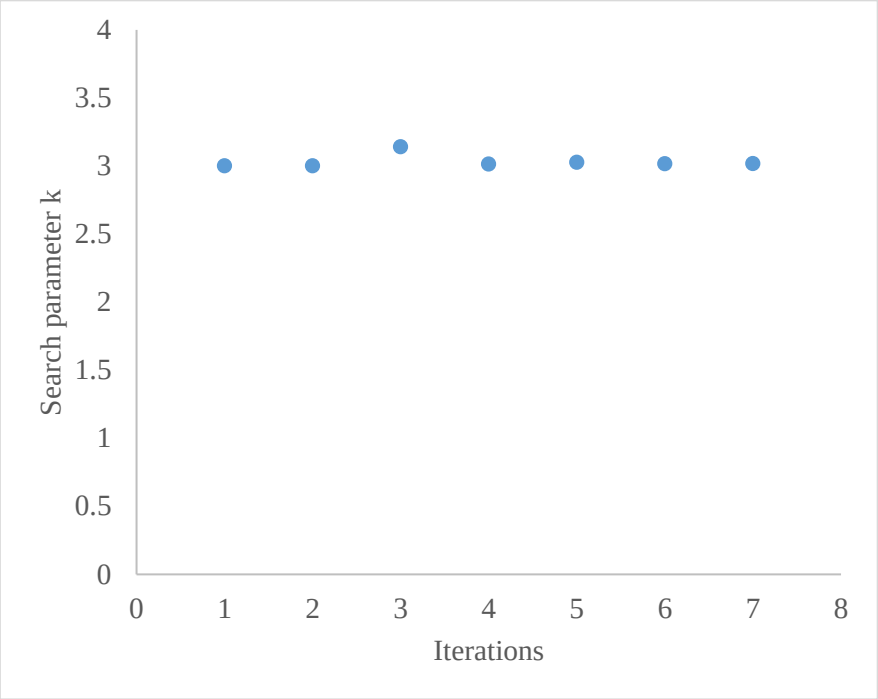

Figure 5.4: Evolution of search parameter k with ARF iterations (case2)

To have a more comprehensive analysis, the distributions of each iteration is propagated. Figure 5.5 shows the model predictive accuracy in terms of RMSE with respect to the number of simulations for both cases. Since, the parameters are well estimated to their true values, Figure 5.5 does not separate between the training and the validation scenarios, however, it plots the model predictive performance in fitting to the whole data comprising the six scenarios.

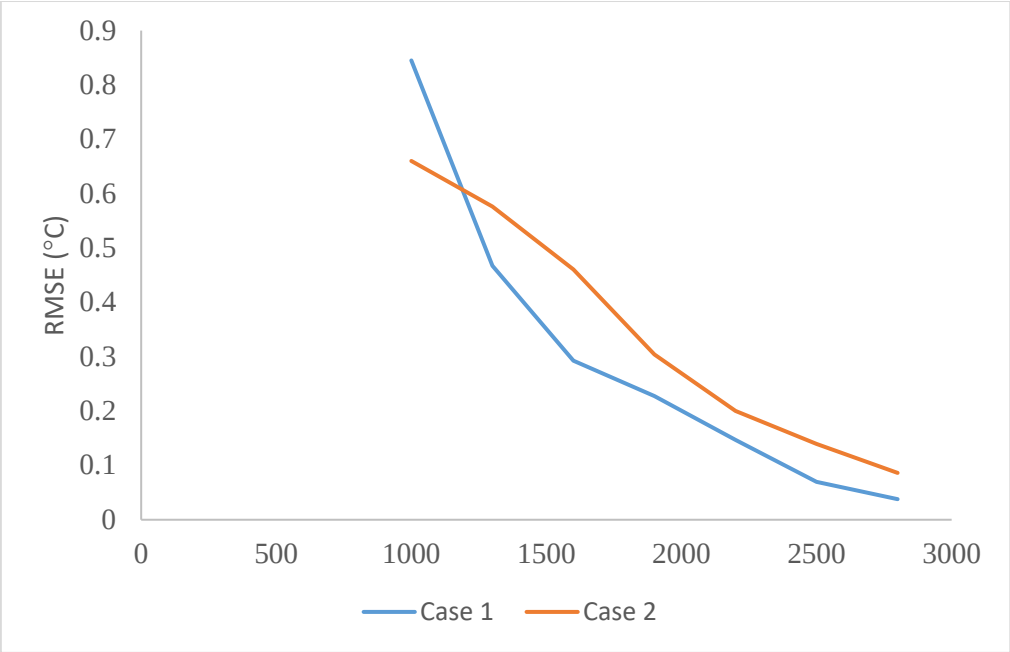

Figure 5.5: Model prediction accuracy (case 1 and case 2)

With a small number of model evaluations, ARF is capable of achieving sufficient accuracy (less than 0.1 °C with 2800 simulation as shown in Fig. 5) in finding the true values and in fitting well to the data. Case 1 reaches a better performance slightly faster than case 2 due to the priors selected: the priors of the 1st case are closer to the true values than case 2. However, it is observed that unexpectedly, at the first iteration, case 2 has a smaller RMSE.

Figure 5.6 and Figure 5.7 show the estimation of the heating power for cases 3 and 4 respectively. In case 3, the confidence region of some priors do not comprise the true values while for other parameters it does. In case 4, all the priors are shifted by $\pm 4\,\sigma$. It clearly shows how ARF is capable of widening the ranges to find the region of high probability. At the first iteration, the boundaries do not contain the true value (1), however, in the subsequent iterations, it is able to encompass it.

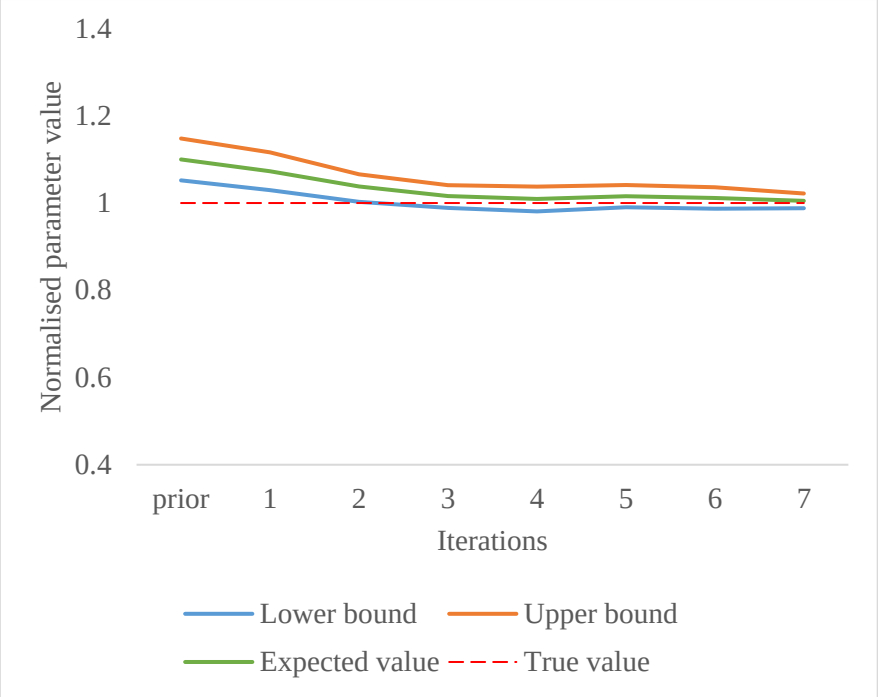

Figure 5.6: Evolution of heating power with ARF iterations (Case 3)

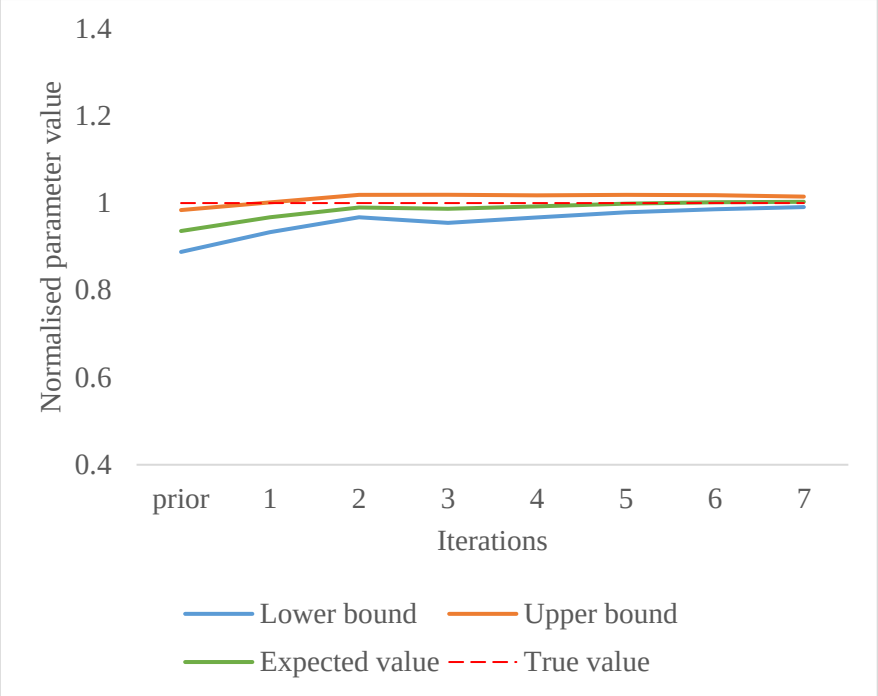

Figure 5.7: Evolution of heating power with ARF iterations (Case 4)

To illustrate the importance of the search parameter, ARF is run on case 3 without adaptively tuning the search parameter k. It is kept at its default value which is 3 at all the iterations. ARF is able to widen the sampling bounds of the heating power correctly in a way that it could easily sample around the true value as shown in Figure 5.8. However, without the

adaptive tuning of the search parameter, ARF is not able to shift all the parameters towards the true value. For example, the solar albedo is stuck in a local minimum below its true value, as shown in Figure 5.9. The y-axis in this figure is also given in terms of ratio: 1 represents the true value of the solar albedo which is 0.35.

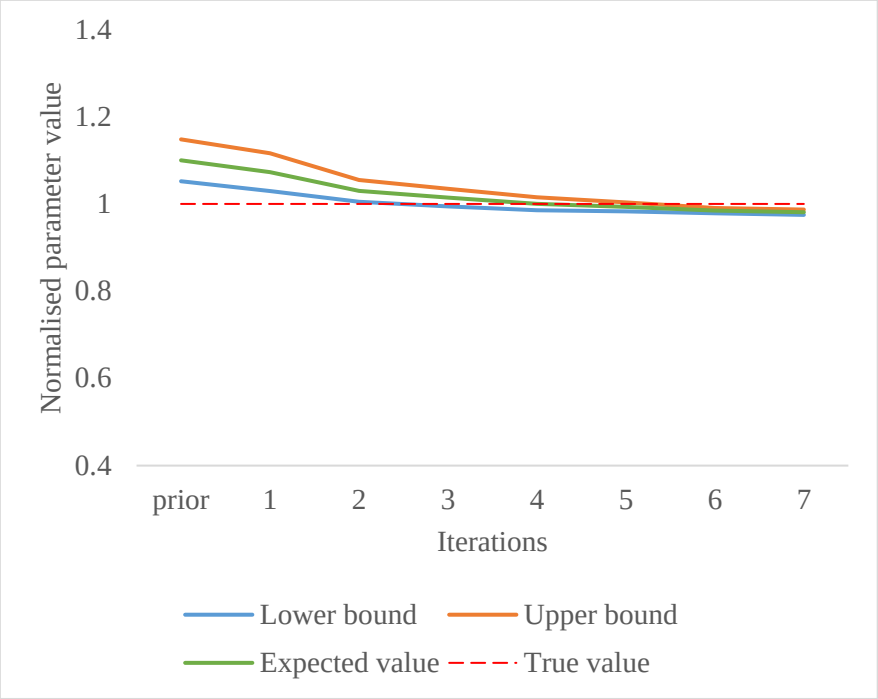

Figure 5.8: Evolution of heating power with ARF iterations without search parameter (Case 3)

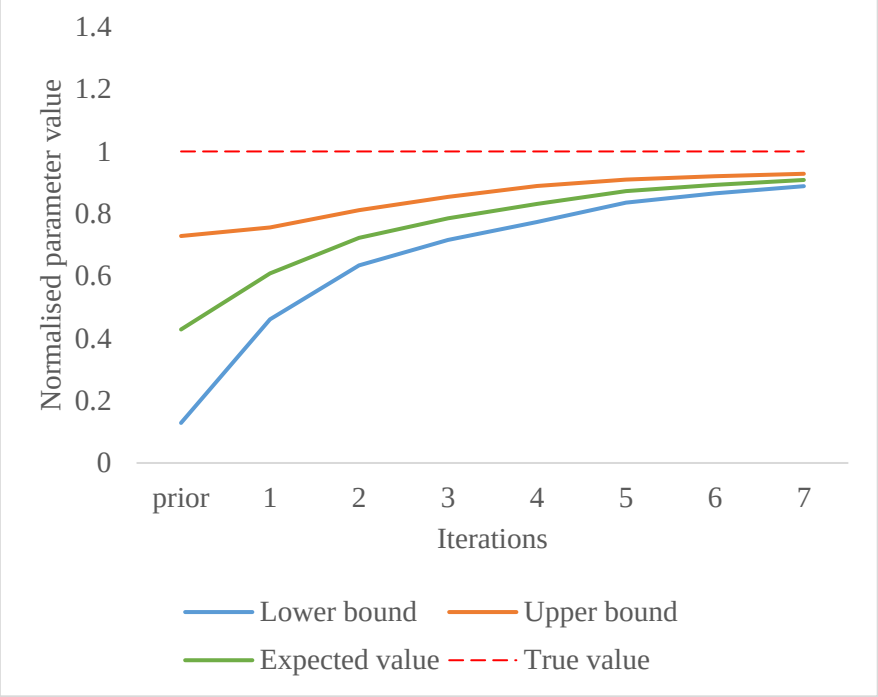

Figure 5.9: Evolution of solar albedo with ARF iterations without search parameter (case3)

This is avoided by adaptively tuning the search parameter. Figure 5.10 shows the estimation of the solar albedo in case 3 when the search parameter is adaptively tuned as is clearly depicted in Figure 5.10. It is shown that there is a significant increase in the upper bound from the third to the fourth iteration.

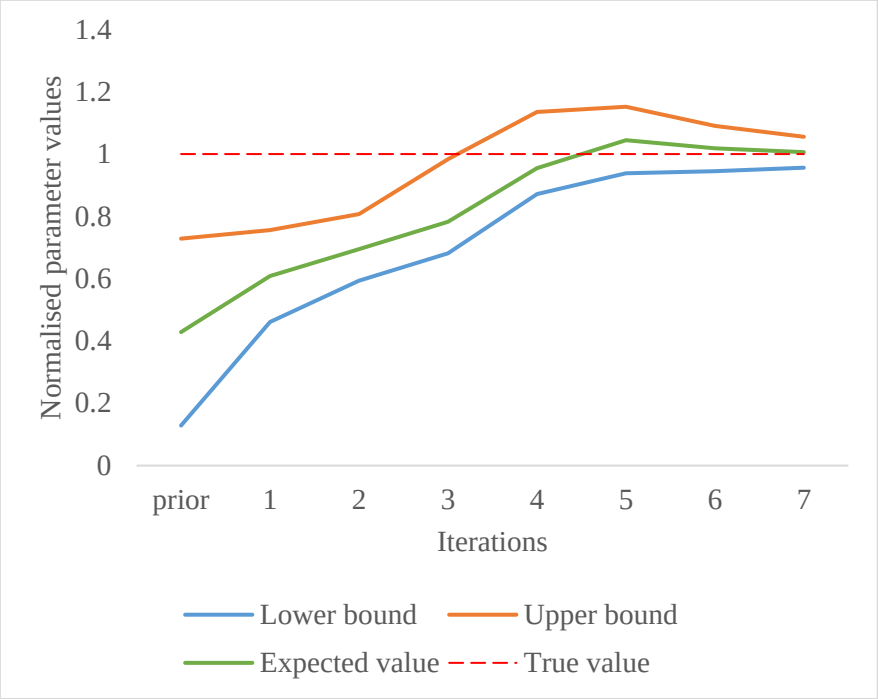

Figure 5.10: Evolution of solar albedo with ARF iterations with search parameter (case3)

Figure 5.11 shows the variation in the search parameter induced on the solar albedo corresponding to each iteration. It shows how the search parameter k widened the search area when necessary in the 3rd, 4th, and 5th iterations and then narrowed it starting from the 6th iteration when the algorithm found the high probability regions around the true value.

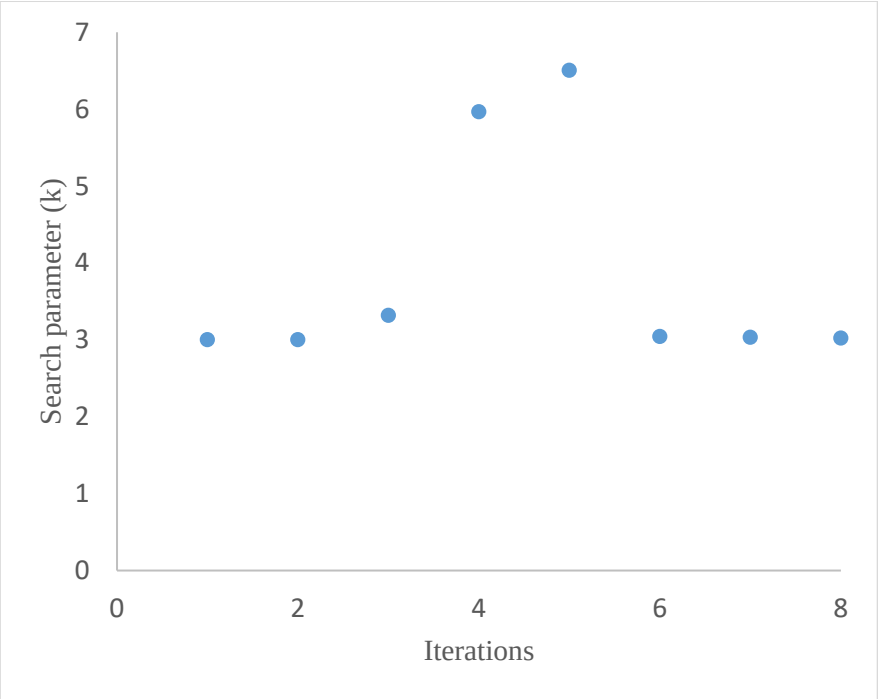

Figure 5.11: Evolution of search parameter k of solar albedo with ARF iterations (case 3)

Figure 5.12 shows the model predictive performance with respect to the number of simulations: the distribution at each iteration is propagated for cases 3 and 4. It illustrates that even with shifted and narrow priors, ARF can overcome these challenges and fit very good to data. Case 4 is found to be slightly faster than case 3. This is expected since most of the priors in case 4 are closer to the true values than case 3. Globally, similar accuracy after 2800 simulations is reached for both cases. This indicates that ARF is also robust against shifted priors.

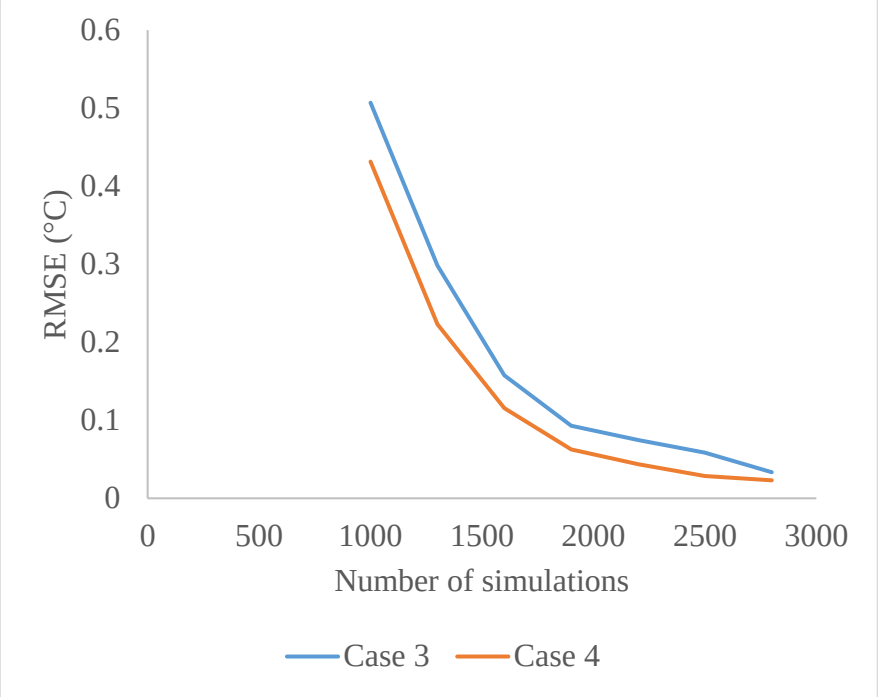

Figure 5.12: Model prediction accuracy (case 3 and case 4)

Case 5 is dedicated to perform a comparative analysis between the algorithms presented in chapter 3 and ARF. APMC is selected among the algorithms presented in literature since it showed the best performance compared to the rest. Moreover, ABC-RF is also retained for the comparison since it is the basis of ARF. A simulation budget of 30,000 is considered for all the algorithms except for ARF, where it is terminated when its RMSE drops below that attained by the other algorithms after 30,000 simulations.

It is hard to fix APMC at a definite simulations budget due to the nature of the sequential sampling adapted. Accordingly, the parameters estimated after an iteration that corresponds to a total simulations close to 30,000 are retained (iteration 96 with a corresponding 29,700 simulations). On the contrary, it is easy to fix ABC-RF at a specified simulation budget which corresponds to the data set size on which the random forest is trained.

Figure 5.13 shows the Euclidean distance between the parameters distributions and their true values obtained by the three algorithms with increasing number of model evaluations. The normalised Euclidean distance between the posteriors and the true values attained by APMC is 0.11. On the other side, closer distributions are estimated by ARF (d=0.082) after only 2200 simulations. With 1000 simulations both ABC-RF and ARF perform similarly and considerably better than APMC. This is expected since the first iteration of ARF is exactly similar to ABC-RF algorithm: ARF differs from ABC-RF in that it incorporates additional iterations.

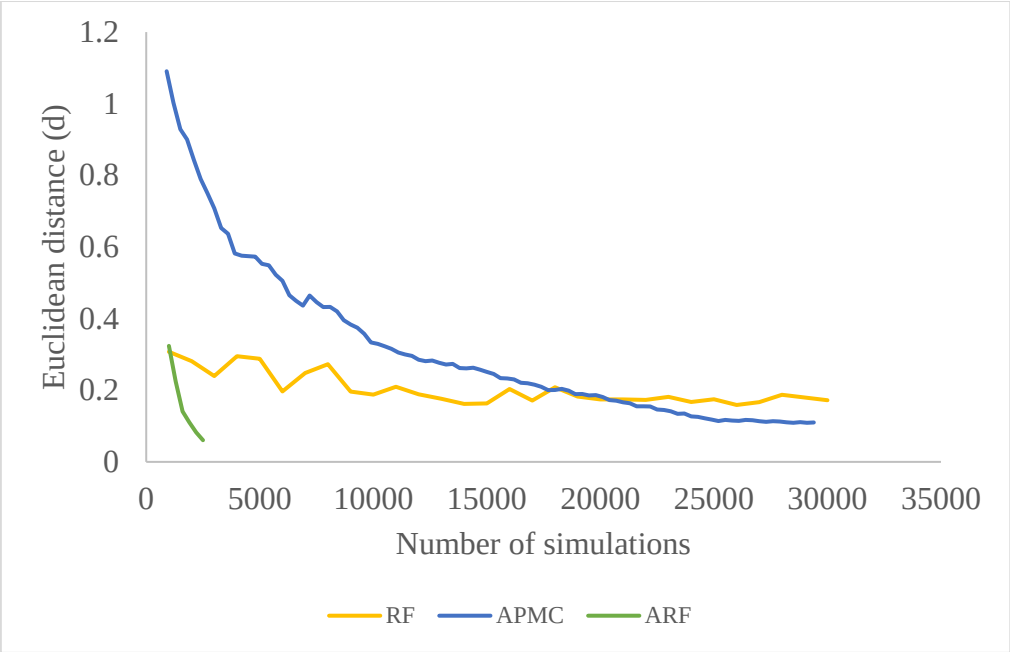

Figure 5.13: Euclidean distance between the parameters distributions and their true values (case 5)

Figure 5.14 shows the model predictive performance with respect to the number of simulations of the three algorithms in case 5. The model predictive performance with APMC decreases continuously to reach an RMSE of 0.036℃ after 29,700 model evaluations. ABC-RF reaches an RMSE of 0.0992℃ after 30,000 simulations which is not the minimum value due to variabilities. ARF is terminated when the accuracy of the parameters distribution became similar to that attained with APMC, that is an RMSE close to 0.036°C. Hence, ARF is stopped at the fifth iteration with an RMSE of 0.031°C corresponding to 2200 model evaluations. This shows that ARF is able to achieve similar accuracies with a considerable reduced number of model evaluations. The method convergence is shown in Fig.5.14 but it is not statistically proven yet. In this evaluation also, the number of trees and size of leaves were not increased according to the number of simulations.

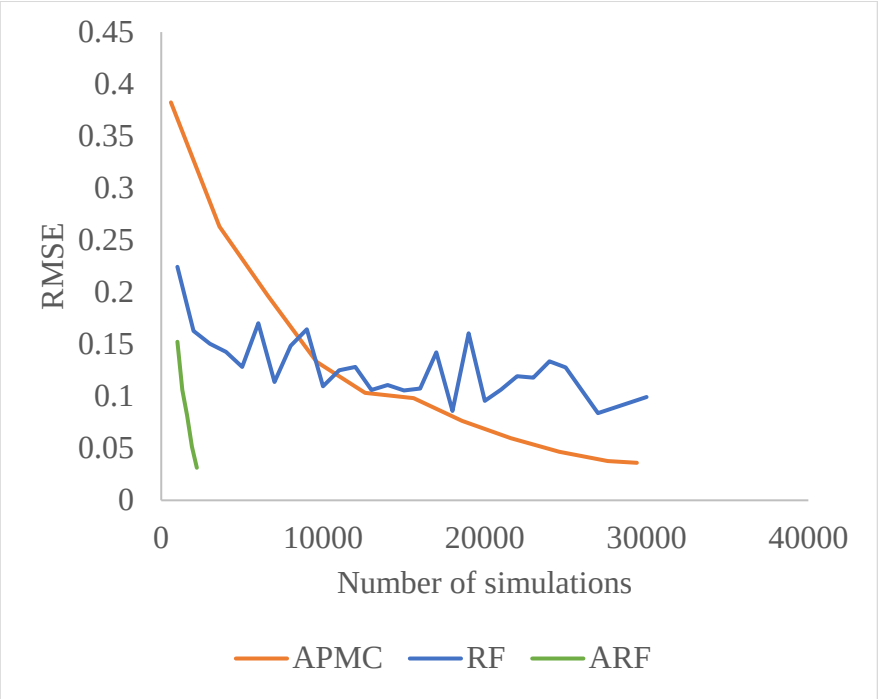

Figure 5.14: Model prediction accuracy (SMC, ABC-RF, ARF) (case 5)

It is shown in chapter 3 that ABC-RF explores well the parameter space but cannot narrow the distributions on the true values. Figure 5.15 to Figure 5.20 show the posteriors attained with ABC-RF and ARF. The posteriors of ABC-RF are those attained after 100,000 simulations. The reason behind going to 100,000 simulations is that random forests generally require a big data set to yield robust and more accurate results. Raynal et al. (2019) mentions that 100,000 is a good choice for ABC-RF and if the degree of variability in its results is found significant, they recommend to increase the size of the data set even more. Therefore, in this work, the data set is increased to 100,000 to make sure that this hyper-parameter of the algorithm is not badly identified. The posteriors of ARF are those attained after 2200 simulations.

The orange lines at the bottom of the graphs represent the rug plot of the data set used to build the random forests in each algorithm. A rug plot is a way to display the distribution of a data set. It projects all the data samples on an axis (here the x-axis). It is very similar to the histogram but it does not combine the samples of the data into separate bins. The samples are thus represented as marks on the x-axis (the orange marks in Figure 5.15 to Figure 5.20.). With ABC-RF, the 100,000-size data set is generated directly from the priors, however, with ARF, the 2200-size data set is generated sequentially, where at each iteration new data are generated from regions closer to the posteriors. This can be visualised in Figure 5.15 to Figure 5.20 by the concentration of more marks around the posteriors. This shows that ARF does not only

explore well the parameter space, but also it does not underestimate the posteriors. Note that in these figures, the true values are those present in Table 5.5. The heating power, internal gains, and ventilation flowrate are all divided by their prior mean values, which are also presented in Table 5.5.

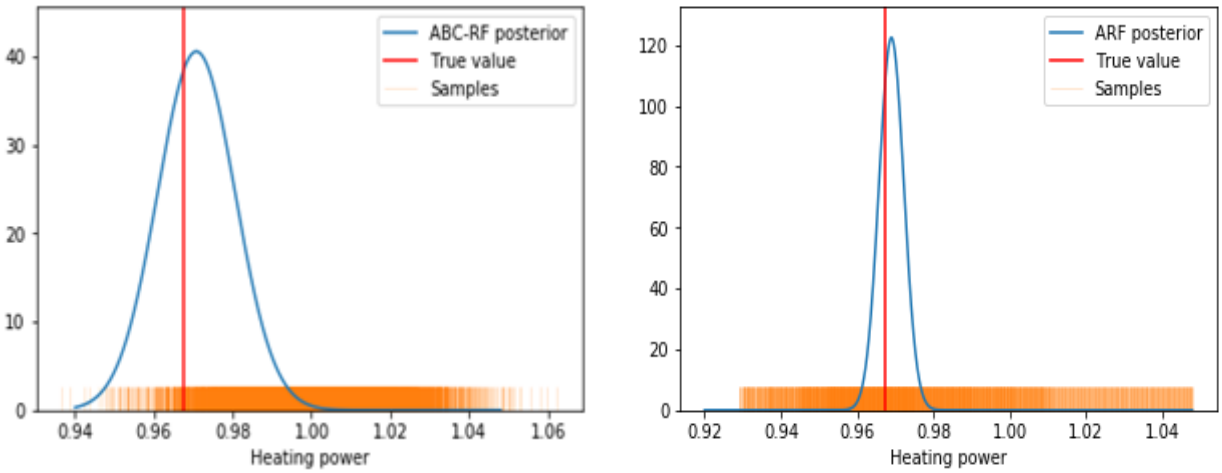

Figure 5.15: Heating power data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

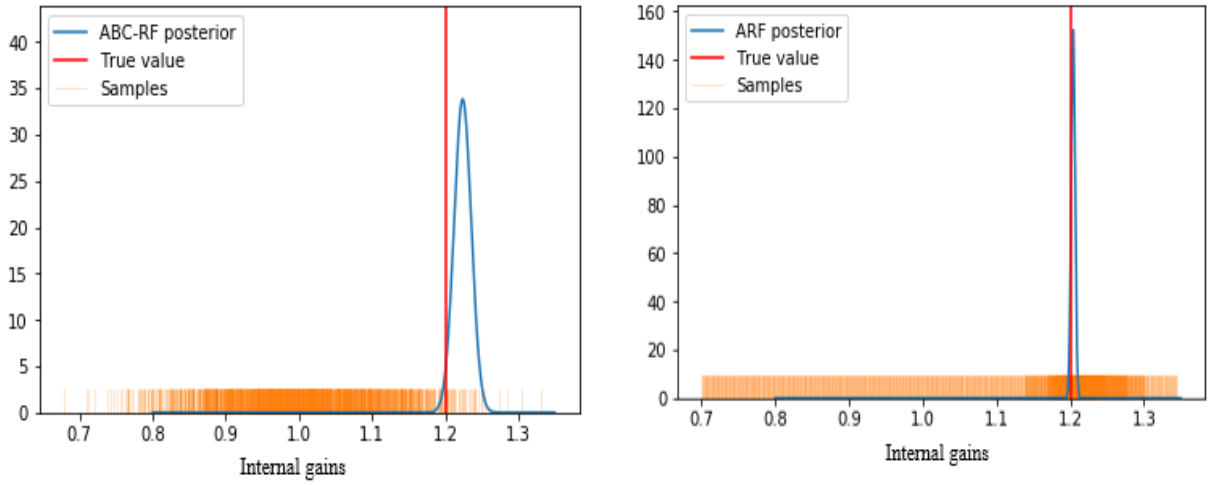

Figure 5.16: Internal gains data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

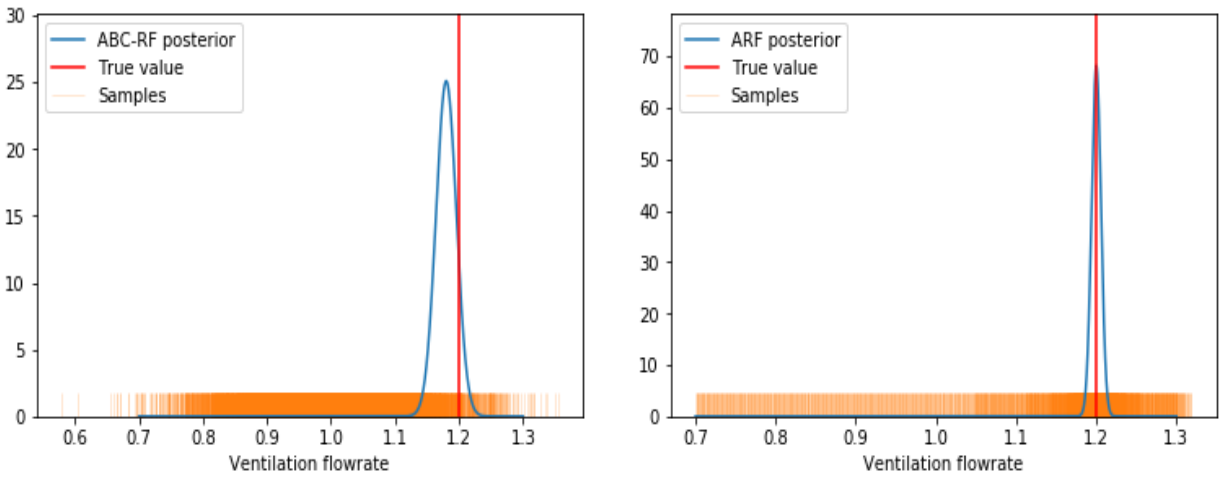

Figure 5.17: Ventilation flowrate data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

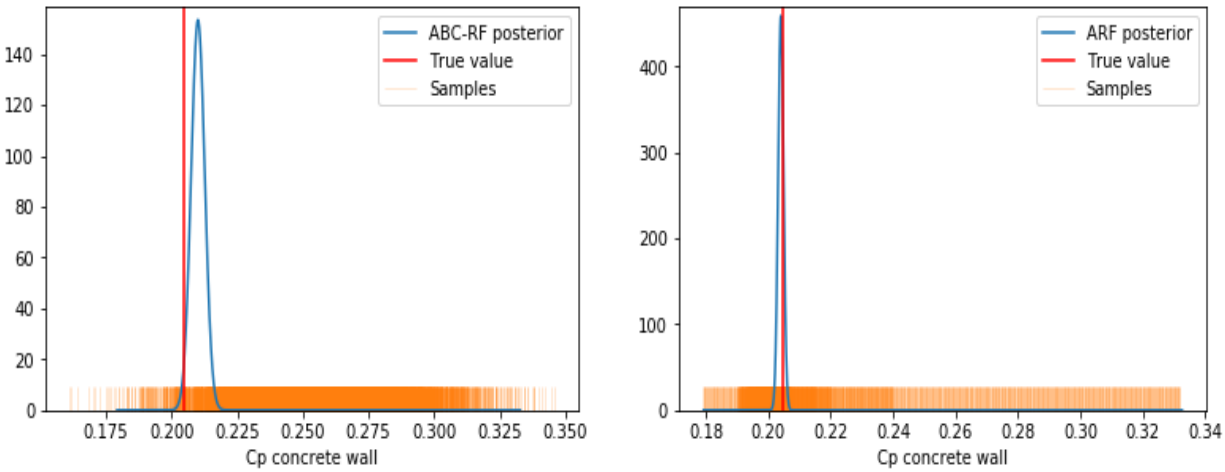

Figure 5.18: Specific heat of concrete wall data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

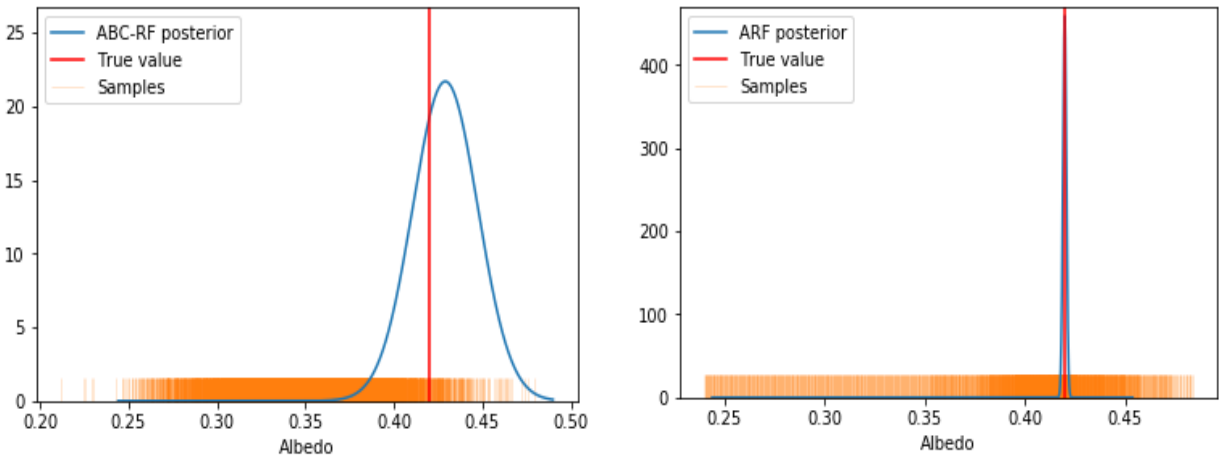

Figure 5.19: Albedo data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

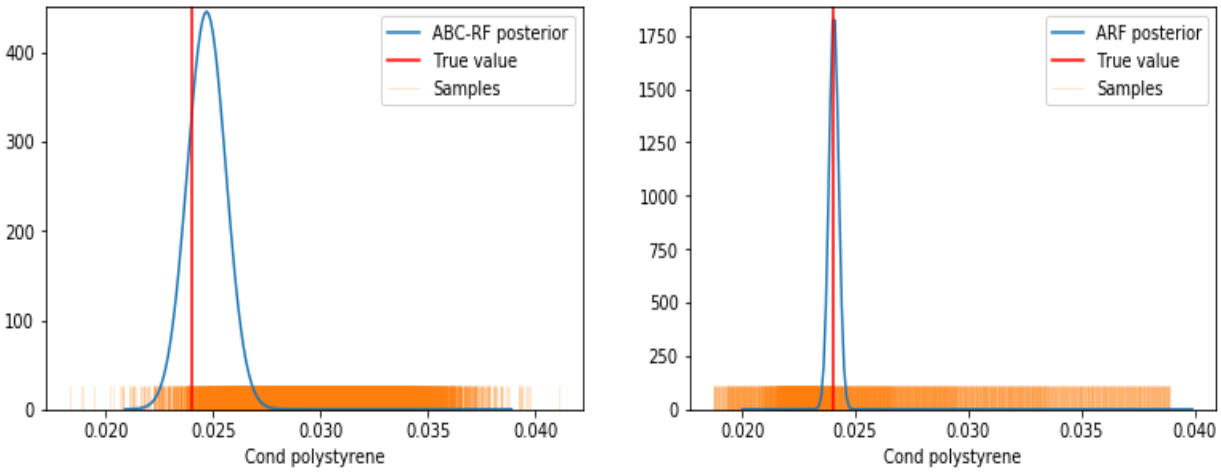

Figure 5.20: Polystyrene conductivity data set and posteriors of ABC-RF (left) and ARF (right) with 100,000 and 2500 samples respectively

The posteriors of ABC-RF attained with a data set size of 100,000 are propagated. The aim is to undergo a comparison in terms of model predictive performance between the original method ABC-RF and its updated version ARF. Figure 5.21 shows the precision of the model prediction in all the scenarios for both algorithms. The number of model evaluations used for these results in ARF is 2500. It shows clearly that ARF can reach a significantly better accuracy than ABC-RF with a lower number of model evaluations.

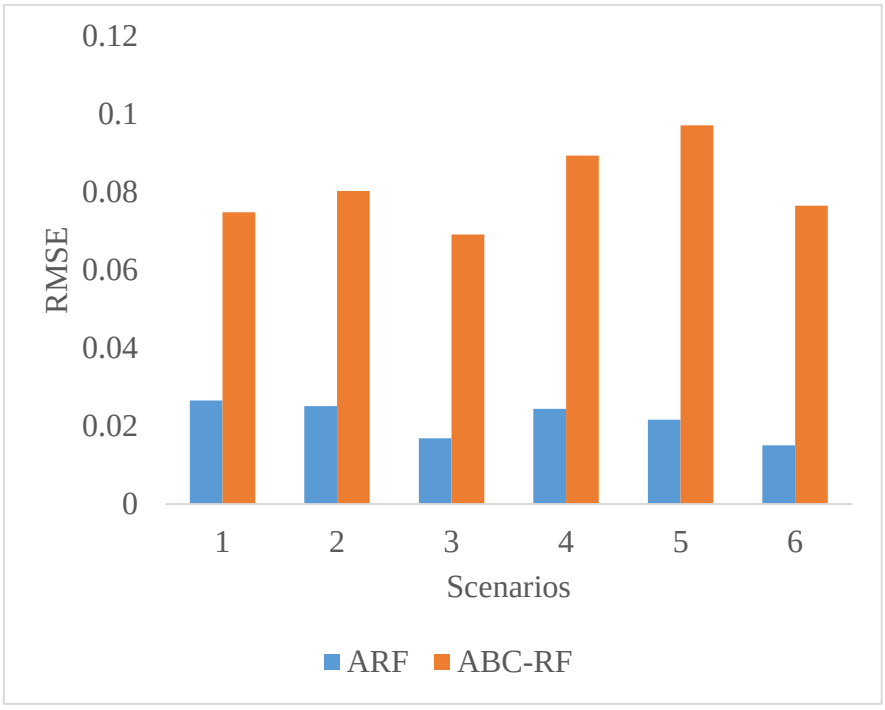

Figure 5.21: Model prediction accuracy for each scenario (case 5)

In all the six cases, ARF is initialised with 1000 samples followed by 300 samples per iteration. Samples could be generated following another form. For example, less samples per iteration results in more iterations which means the parameters space from which samples are generated is updated more often. This more frequent update of the parameter space allows sampling more from the posteriors regions. This might decrease the global model evaluations required. Another option that could also be promising is to follow a decreasing sequence: large sample size at first iterations account for the relatively large variances and then smaller sample sizes are generated. This is one of the algorithm hyper-parameters whose effect on the global performance would deserve further study.

5.6 Application using in-situ measurements

So far, virtual case studies were considered in the previous section. In this chapter, a real case is considered. The objective is to apply the calibration methodology on real in-situ measurements. The same case study illustrated in the chapter 2 is retained but the real monitored temperature profile is considered.

In section 5.6.1, the calibration methodology is presented. This methodology is formulated based on previous literature and on some findings of this thesis. In section 5.6.2, a brief description of the case study is provided. The parameters quantification of the case study are taken from previous work and they are briefly presented in section 5.6.3. In section 5.6.4, all the calibration methods presented in this thesis are applied to the real case study and compared.

5.6.1 Calibration methodology

In this section, a workflow for applying Bayesian calibration to building energy models is proposed based on the findings of this thesis.

One of the major aspects of calibration, which is critical for its performance, is the identifiability of the parameters. To this end, structural identifiability could be applied to simple linear models to detect unidentifiable parameters. These methods could be limited to such simple models. For complex building energy models, sensitivity-based identifiability analysis is preferred since it does not require entering to the model structure. In this thesis, orthogonalisation method is found to perform well in ranking the parameters in terms of

estimability. Accordingly, sensitivity analysis preceded by orthogonalisation is advised to be applied as a basis for parameter selection.

Another critical choice for calibration is the number of parameters that are included. To this end, based on two virtual case studies applied in this thesis, it was shown that the parameters ranked with an estimability threshold of 0.04 are identifiable. It is also shown that similar performance in terms of model predictive performance with lower thresholds. Accordingly, the selection of this value is still arbitrary. For the following case study, a cut-off value of 0.04 is considered to separate the estimable parameters from the non-estimable ones.

The five calibration methods reviewed in chapter 3 along with the ARF method proposed in this thesis are applied to the real case study. The indicators used in chapter 3 are retained to perform this comparison.

5.6.2 Case study

The same house described in chapter 2 is retained. The meteorological data correspond to the measured values (outdoor temperature, global solar radiation and horizontal diffuse radiation) at hourly time intervals at Le Bourget-du Lac airport (France) located near the I-BB house: the shading induced by the surrounding mountains are taken into account in the measured data. The information relating to the site is:

- Longitude: 5.8814°E
- Latitude: 45.6876°N
- Altitude: 233 m
- Average ground temperature at 10 m depth: 9°C.

For the experimental campaign, eight temperature sensors are mounted at a height of 1.10 m in the different rooms. The sensors are platinum probes protected by a heat shield to prevent radiation from influencing the measurement. In this study, a single thermal zone is considered. Consequently, the interior temperatures measured in the rooms of the house are weighted in proportion to the net floor area of these rooms in order to compare their average with the interior temperature simulated by the monozone model.

Similarly to chapter 2, the different scenarios of the experimental campaign are split into training and testing. Scenarios 1, 2, 3 and 5 are taken as training scenarios and scenarios 4 and 6 are kept for validation.

5.6.3 Parameters quantification

Munaretto (2014) conducted a detailed parameters quantification for this case study. Here, the quantification of some parameters that are considered in the subsequent sections is provided.

The house is equipped with a double flow controlled mechanical ventilation (CMV). Fresh air is blown into the living room and the bedrooms (flowrate V1), while outlets located in the toilets and the bathroom ensure the extraction of stale air (flowrate V2). After in-situ measurements conducted by a French research institute (Commissariat à l'énergie atomique et aux énergies, CEA), the nominal ventilation flowrates $\dot{V}_1$ and $\dot{V}_2$ are set at 110 and 160 m^3/hr respectively. The maximum heating power during the experimental campaign is 1200 kW. Over the period of the study, the temperature setpoint is never reached, so the electric resistance is working all the time at its maximal power. The total internal gains in all the different rooms generated by different equipment (transformers, measurement processing unit, various sensors, CMV motor) is measured to be 208 W. The house is surrounded by white sand and grass. An albedo-meter is set up to measure the average reflexivity of the surrounding ground and a value of 0.35 is measured.

5.6.4 Application and results

The sensitivity and identifiability analysis are taken from the previous chapter and the same ranking is considered here. The six most estimable parameters which are in this case also the six most influential parameters are calibrated using the real data. The prior distribution fit to each parameter are taken from Robillart (2015) except the solar albedo which was modeled as a uniform distribution bounded by [0.28, 0.42].To avoid overfitting towards one of the boundaries, the solar albedo is modeled in this chapter as a normal distribution as illustrated in Table 5.6.

Table 5.6: Prior distributions (Robillart 2015)

Parameters	Distribution	Mean	σ	Unit
Ventilation flowrate	Normal	110	11	$[m^3/h]$
Internal gains	Normal	208	20.8	$[W]$
Heating power	Normal	1200	20	$[W]$
Specific heat of concrete	Normal	0.256	0.0256	$[Wh/(Kg.K)]$
Solar albedo	Normal	0.35	0.035	$[-]$
Conductivity of polystyrene	Normal	0.03	0.003	$[W/(m.K)]$

The five calibration methods presented in chapter 3 are firstly applied and compared based on their predictive performance. Figure 5.22 shows the RMSE of the posteriors and of the priors after a simulation budget of around 44,000 model evaluations and Figure 5.23 shows the temperature profile in all the scenarios. All algorithms yielded a more precise model to adequately fit the data. There is a significant decrease in the RMSE value between the calibrated and the un-calibrated model, with APMC being the best and ABC-RF being the least accurate. The performance is good on the training scenarios (1, 2, 3 and 5) and on the testing scenarios (4 and 6) even though, some scenarios are better fitted than the others.

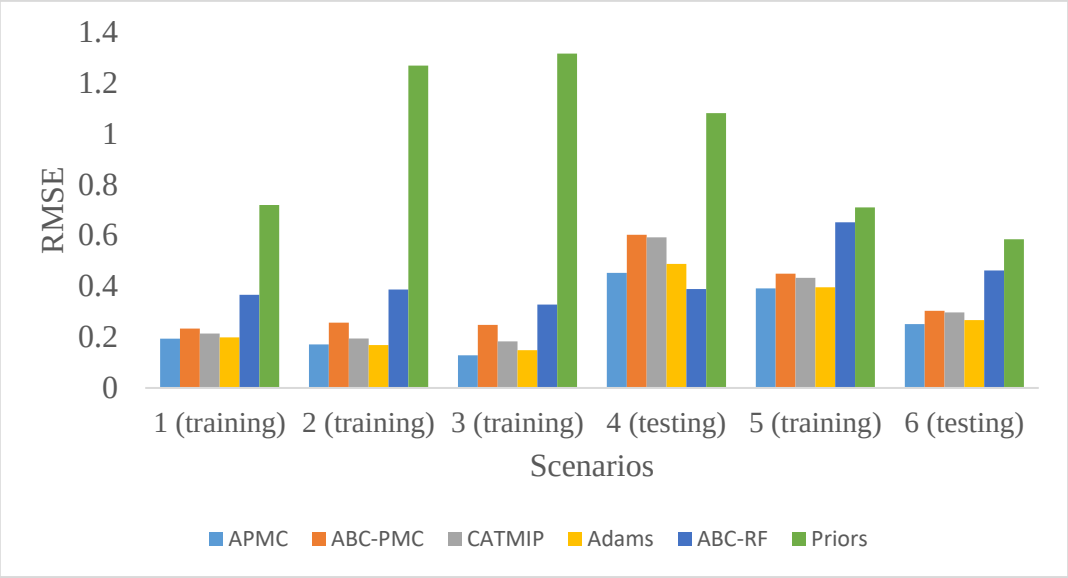

Figure 5.22: Model prediction accuracy of calibration algorithms

Robillart (2015) who proposed using the sequential ABC algorithms to calibrate building energy models applied an ABC algorithm on the same case study. However, the algorithm applied did not adaptively compute the thresholds at each iteration, instead the sequence of decreasing thresholds need to be identified beforehand to initialise the algorithm. The

calibration performed was able to obtain better posteriors than priors. The propagation yielded an RMSE 0.8 and 0.5 on the testing scenarios 4 and 6 respectively. In this work, using the improved algorithms, the model predictive performance of the calibrated model was enhanced even more to reach RMSE of 0.45 and 0.25 on the same two scenarios respectively.

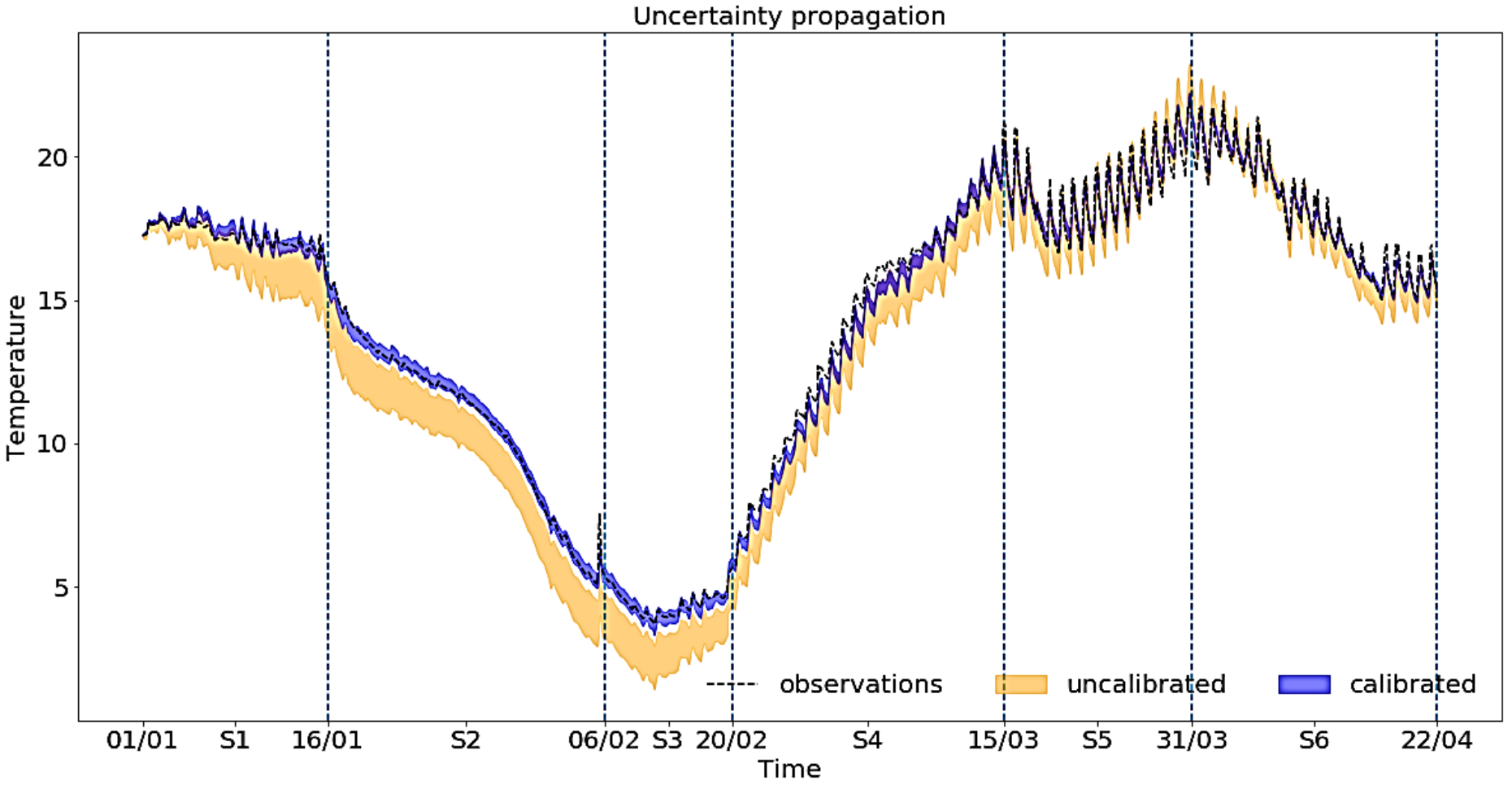

Figure 5.23: Uncertainty propagation on temperature profile (S1: scenario1; S2: scnenario2; S3: scenario3; S4: scenario4; S5:scenario5; S6: scenario6)

The accuracy of the algorithms is consistent with what was observed on the virtual data. Figure 5.24 shows the performance during the evolution of the algorithms. In this case, unlike previously, its slope is decreasing with increasing simulations and it seems like the methods are converging to an RMSE higher than the threshold. This means that the model is having difficulty in fitting the actual behaviour better. This could be related to different issues such as measurement bias and uncertainty, error in the specification of the un-calibrated parameters, or errors due to model assumptions. It can also be emphasised that the algorithms show similar relative computational efficiency as depicted previously on the virtual data.

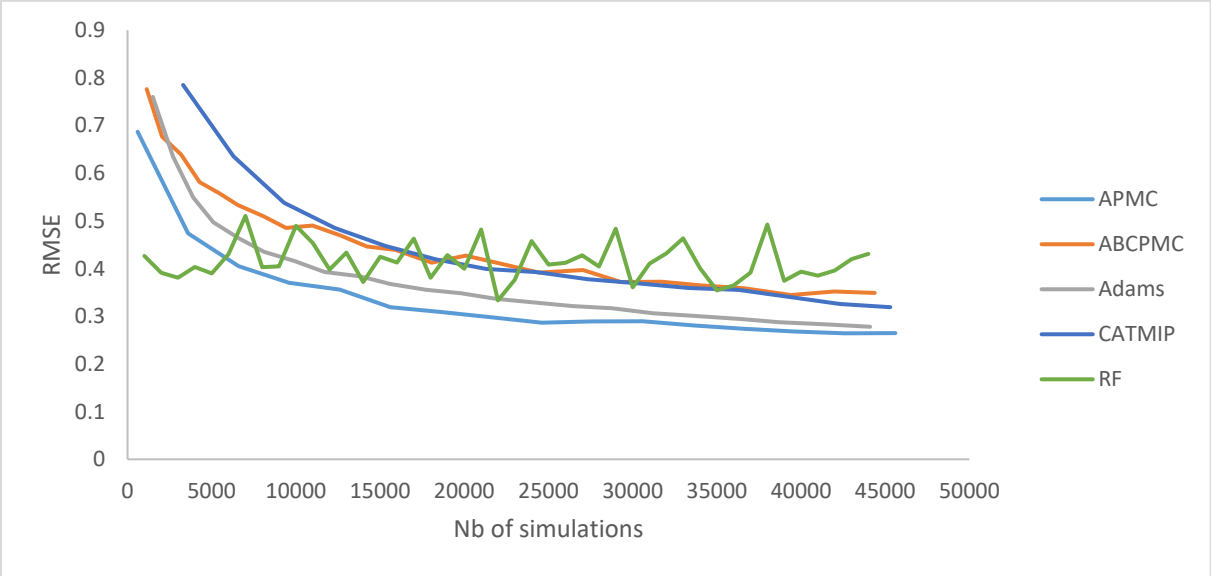

Figure 5.24: Model prediction accuracy evolution (real measurements)

ABC-RF was also run with a data set size of 100,000 samples and there were no major differences with the results obtained with a smaller size under default hyper-parameters values.

One of the main aspects of ABC-RF is that it cannot predict values outside the range of the priors. Thus, if the true value of the process lies outside the boundaries of the priors, ABC-RF could only favour the samples that lie in the prior and are closer to the true value. That is to say that ABC-RF is less robust to the selection of the priors than the other algorithms. To clarify this finding, Figure 5.25 depicts the prior and the posterior of the heating power obtained with APMC and ABC-RF algorithms. We note that the posterior is quite similar for all algorithms except for ABC-RF.

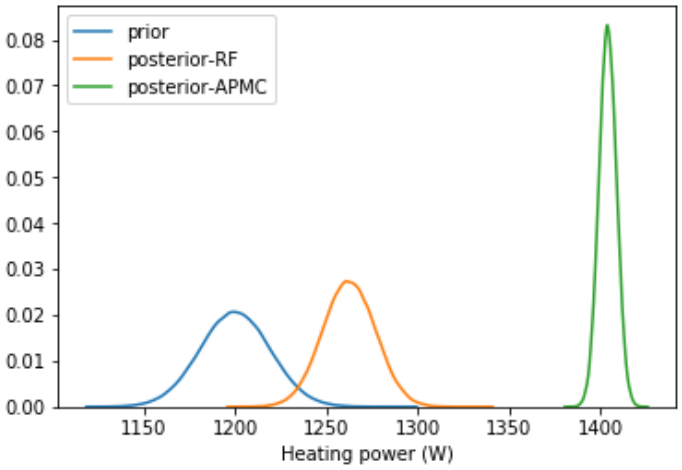

Figure 5.25: Prior vs posterior (heating power) APMC, ABC-RF

The mean of the posterior obtained with ABC-RF is 1262 W, whereas, for the other algorithms, it is around 1400 W. The same behaviour is also noticed with the specific heat of the concrete wall. To improve the performance of this algorithm, it could be better to increase the ranges of the priors which would require a larger data set to explore well the parameter space. However, since it is observed that ABC-RF can yield, with a small data set, similar results as with a big data set, it could be better to increase the prior range with a similar data set size.

Figure 5.26 displays the posterior against the prior of the APMC algorithm corresponding to the specific heat of the concrete. Similar posteriors are obtained by the other algorithms with slight differences.

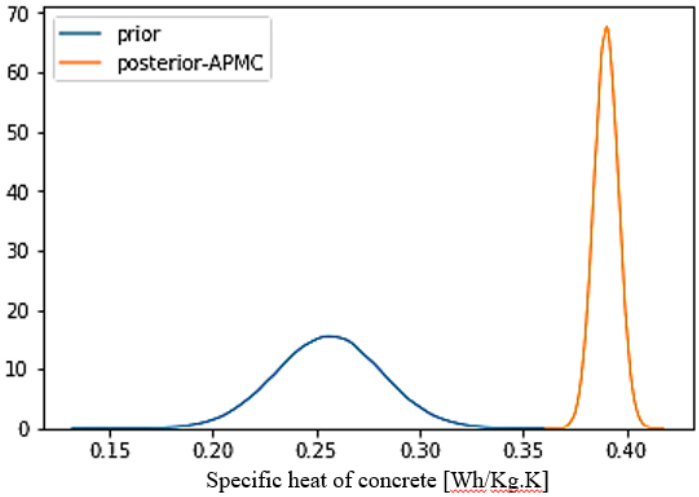

Figure 5.26: Prior vs posterior of specific heat of concrete (APMC)

It is clear that the posterior distribution overestimates the real value of the specific heat. In fact, the prior of this parameter is determined with a high degree of confidence because it is based on the type of concrete and since this house is constructed for laboratory experiments. So it is surprising to see such shift that is quire large. Only the six most influential parameters are calibrated, however, it was shown in the previous chapters that the specific heats of concrete screed and slab are also influential, but they are not considered in calibration. This means that the global underestimation of the global thermal capacity might be compensated by the specific heat of concrete wall. However, the calibrated model is able to perform well even on experimental data. If the objective is to have a well calibrated model that fits to real measurement, this is an acceptable behaviour. However, this method must be used with caution to identify the value of a specific parameter.

Another point regarding the precision attained for the different scenarios should also be highlighted. In Figure 5.22, it is observed that unlike scenarios 1, 2 and 3, the RMSE values for scenario 5 are relatively large. The model was calibrated on these four scenarios; scenarios 4 and 6 were left for validation. The reason behind this difference can be related to the fact that the building was modelled as one thermal zone, and that the measured temperature profiles of rooms were averaged. Figure 5.27 adapted from Munaretto (2014), illustrates the temperature evolution of the different zones of the building. In the first three scenarios, all the zones had a quite similar temperature evolution. However in the 5th scenario, the variability caused by the exterior temperature on each zone is different. For example, the variability in bedroom 3 (Ch.3) is less than that in the living room (Salon). A monozone model could not replicate these different variabilities in the zones.

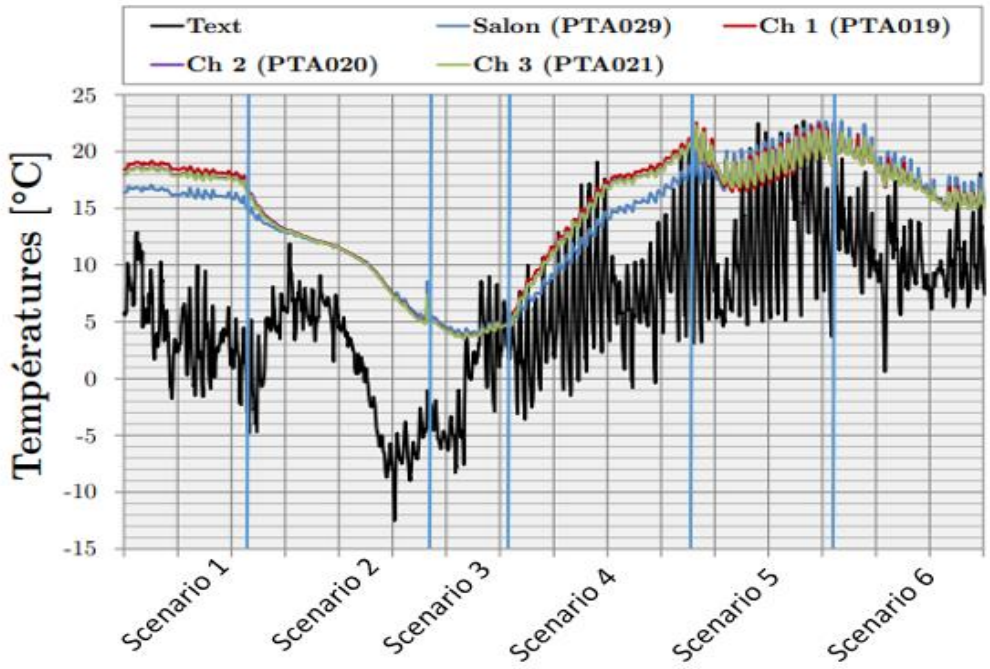

Figure 5.27: Temperature profile for different zones (Munaretto, 2014)

To verify this analysis, the temperature profile of the posterior model simulation was plotted against the measured temperature evolution (the average of all the rooms) as shown in Figure 5.28. It is clear how the monozone model could not precisely replicate the variability shown in the measurements.

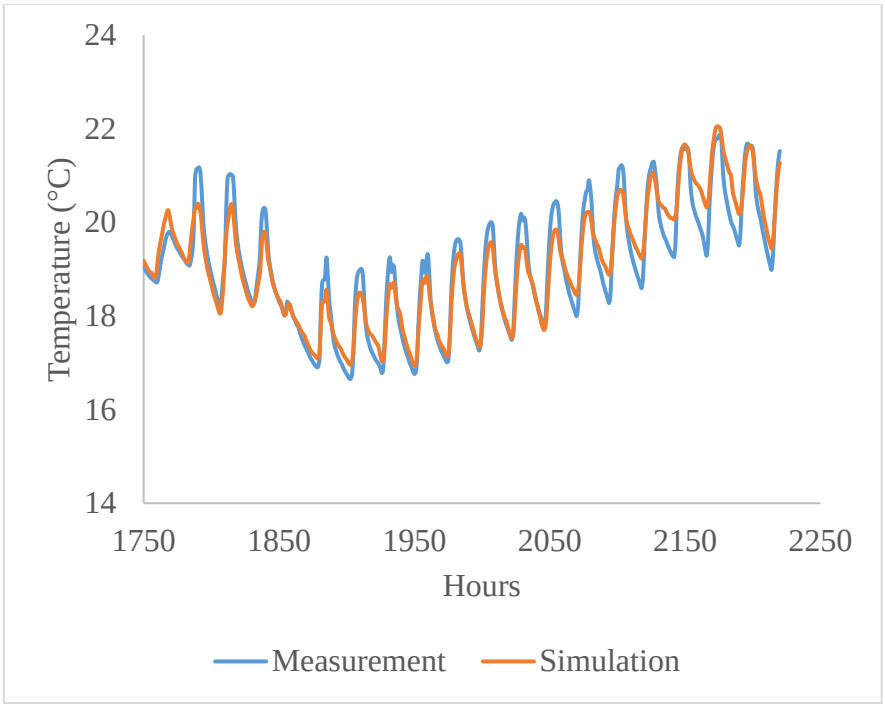

Figure 5.28: Simulation vs measurement temperature evolution (scenario 5)

A multizone model of the building needs to be established to better approximate the real behaviour with as little bias as possible, especially in this case study which is intended for scientific research and has relatively low uncertainty. This application allows to verify if the monozone assumption and thus faster model was a good assumption or not.

The proposed method ARF was also applied on the same case study. Contrary to the behaviour depicted on the virtual data, the model predictive performance of ARF does not show a continuous increase in the precision with more iterations. Figure 5.29 shows the RMSE averaged over all the scenarios as a function of the algorithm iterations.

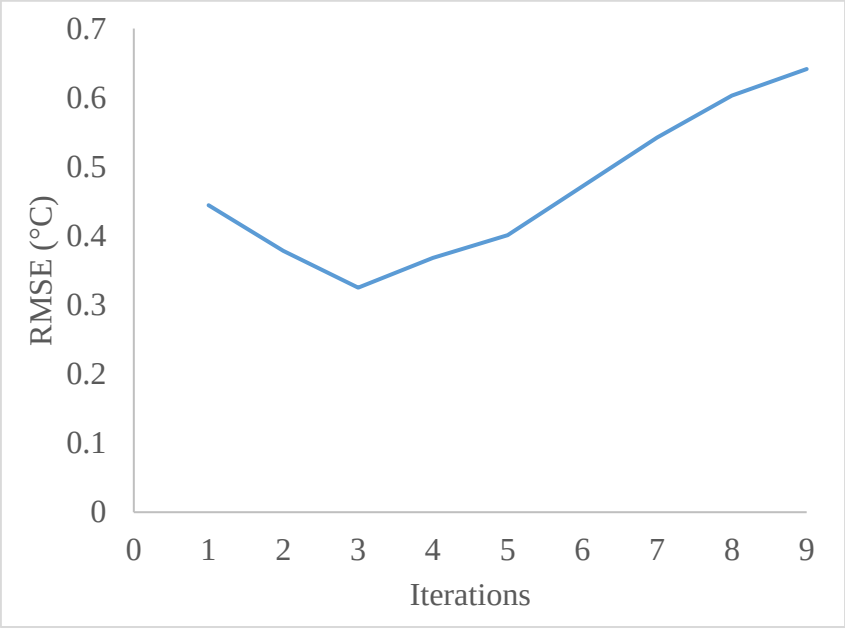

Figure 5.29: Model prediction accuracy for ARF (real measurements)

It is clearly depicted that the minimum RMSE obtained corresponds to the third iteration, which is reached with a 1600 model evaluations. After the third iteration, the RMSE started increasing. This behaviour is different from what is observed for the other algorithms applied in this chapter. To compare the algorithms, the posteriors of the third iteration is propagated and compared to the posteriors of the other algorithms attained with a simulation budget of 44,000 (Figure 5.30). ARF shows a better performance than RF. Compared to CATMIP and ABC-PMC, ARF shows an approximately similar performance, and compared to APMC and Adams, its performance is slightly worse.

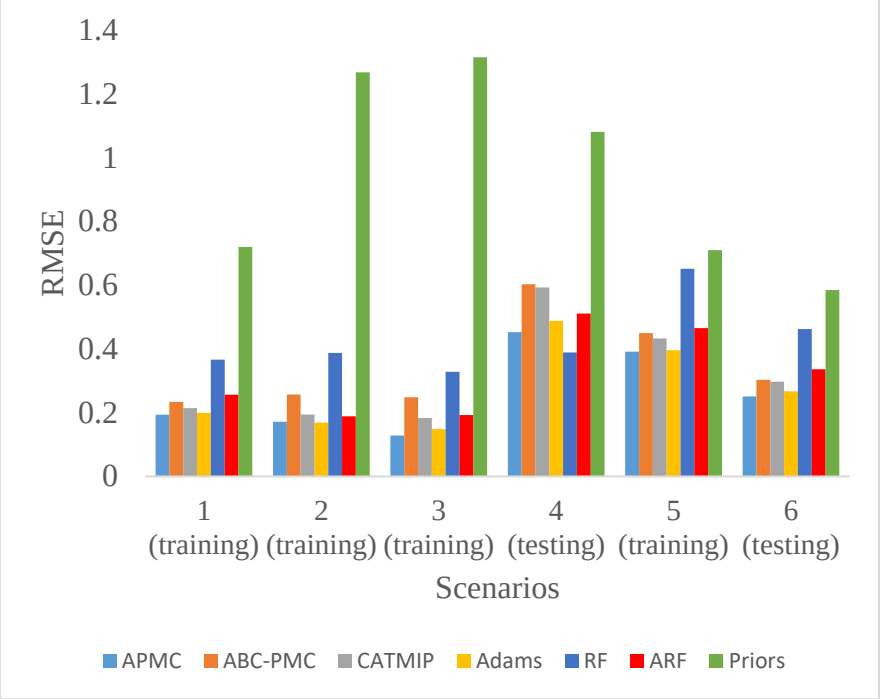

Figure 5.30: Model predictive performance of the algorithms (ARF: 1600 simulations, rest: 44,000)

It is important to mention that the comparison between ARF and the other algorithms is held with different number of model evaluations (1600 for ARF and 44,000 for the others). With more simulations, all the algorithms are capable to attain better performance than ARF except RF. However, ARF is observed to diverge with more simulations. The average RMSE attained by ARF with 1600 simulations is 0.33 ℃. Table 5.7 shows for each algorithm the required number of simulations to reach this value of RMSE.

Table 5.7: Simulation required by calibration algorithms to reach ARF minimum RMSE

Algorithm	Number of simulations
APMC	15600
Adams	21900
CATMIP	42300
ABC-PMC	49065

One advantage of ARF is that it explores well the whole parameter space very efficiently; however, even if good distributions are found at a given iteration, it might shift away from them towards worse distributions in subsequent iterations. In ARF framework, at each iteration, all the newly simulated samples are retained to train a new random forest. If the new samples are all worse than the previous ones, the trained random forest could yield worse results than in previous iterations as observed in this case study. One way to deal with this is that instead of retaining all the simulated samples at a given iteration, only those that yield RMSE values less

than the α-qunatile of the RMSE in the previous iteration are retained to run a new random forest. This ensures that when the algorithm proceeds from one iteration to another, only better samples are kept and this could avoid the diverging behaviour observed.

5.7 Conclusion

Calibration of building energy models has recently attracted the focus of researchers in the field especially the application of Bayesian approaches. A significant work is conducted on these approaches to enhance their performance in terms of accuracy, robustness, and computational efficiency. In this chapter, recent developments in the field of Bayesian calibration were used and a new algorithm called adaptive random forest ARF was proposed.

ARF was compared to ABC-RF and APMC. APMC is selected since it showed the best performance compared to the other algorithms presented in chapter 3. ABC-RF was also retained for comparison since it is the original method on which ARF is based. ARF overcomes the variabilities present in ABC-RF and the computational burden present within APMC. With only a few thousand model evaluations, ARF is capable of attaining accurate results which is considered very fast compared to other methods in the field (it is shown to be more than 10 times faster than some methods). This computational efficiency replaces the need of using a metamodel, which poses an additional problem of error and uncertainty to the calibration.

ARF is also able to overcome the extrapolation problem with ABC-RF. It can easily sample outside the ranges of the priors without requiring considerable additional model evaluations and iterations. This means that ARF is robust relative to a poor prior definition; however, further robustness analysis should be conducted on other case studies. Another very interesting aspect concerning ARF is that it is robust (repeating calibration leads to similar results), but its convergence is not statistically proven though it is shown empirically in the case study.

ARF is tested on temperature profile as the calibration data. It would be interesting to analyse its performance on the hourly heat consumption, monthly, and yearly heat consumption. ARF is terminated when its RMSE dropped below the ones reached by other methods. This is sufficient to perform the comparison between the methods, however, it is also important to simulate more iterations to analyse better its performance.

The various methods that are presented in chapter 3 are applied to the case study with in-situ measurements of the temperature profile. Except for ABC-RF which does not perform very well, the parameters estimation are quite similar for all the algorithms with slight differences while maintaining similar model predictive performance. Globally, all the considered algorithms enhanced the performance and the capability of the model to fit to the training and the testing data. Further improvements could have been attained if the building under investigation had been simulated based on a multizone model instead of a single thermal zone. Moreover, the training and testing scenarios can be swapped. This allows to assess better the case study and to verify the sources of uncertainties. It will also allow to have more testing scenarios on which calibration can be validated. For instance, calibration can be performed on one or two scenarios, which keeps four testing scenarios. This can also be used to assess how the calibration is affected by the quantity and quality of the data.

ARF was also applied with real in-situ measurements. It was observed with ARF, that on real data a divergence in the model predictive performance occurred with more iterations. At the third iteration with 1600 simulations, it yielded the best posteriors. These posteriors were compared to the posteriors of the other algorithms. The comparison shows that for a limited simulation budget of 1600, ARF performed the best in fitting to the data, however, with more simulations, the other algorithms yielded a better data fit.

With the controlled conditions, the algorithms performed better than with real measurements. Multiple factors are probably involved. Firstly, the uncertainties in the influential parameters that are excluded might be presumed into the parameters considered and caused such problem. The bias in the model, measurements and the noise also have their influence on the parameters estimation.

The behaviour of ARF was different to what was observed with virtual data. The is related to the unknown uncertainty and error in the model and measurements. The reason behind the continuous decrease in the RMSE with iterations in the other algorithms compared to ARF is the sampler. In ABC-PMC and APMC, only better samples are accepted at each iteration. Likewise with Adams and CATMIP, due to the incorporation of the Metropolis Hastings method of MCMC, the sampler only occasionally accepts samples that are less probable than a current sample. In ARF, all the samples generated at a new iteration are used without comparing their probability with the samples of the previous iteration. This could be modified by only retaining the samples with higher probability of occurrence than those in the previous iteration.

This might solve the problem of divergence, but on the other side it is expected that the total computational cost of the algorithm might increase. Another option would be to increase the number of trees and leaves size according to the number of simulations. Moreover, it is important to statistically prove the convergence of the method in addition to showing its empirical convergence.

At the moment, without further modification of ARF, one could simply retain the results of the best iteration (with minimal RMSE). This is a good solution if speed is of a main importance, otherwise, if it is focused on precision with less limitations on the speed of calibration, then APMC would be the best algorithm to use among the ones used on this thesis.

General conclusion and perspectives

Due to the fact that the building sector is responsible for the largest share of energy consumption in France and in Europe, researchers are increasingly interested in improving the energy efficiency of buildings. Many countries have set some guidelines and policies to abide by while constructing new buildings to ensure that they comply with the objectives of energy efficient buildings. However, the percentage of newly constructed buildings to those already existing is small. This means that there is a large interest in renovating these existing buildings. Building energy models are normally used to help quantifying and comparing different measures and their possible energy savings. Such an approach based on models is subjected to many sources of uncertainty. This means that these models can allow decision-makings for renovation measures; however, risk management or confidence quantification requires supplementary efforts. To tackle this issue, calibration of building energy models improves the precision of these models in representing the actual behaviour of the building under study. Moreover, Bayesian calibration is an efficient approach in quantifying the uncertainties in the model parameters and the corresponding model predictions of the proposed renovation measures, which allows to represent the predicted energy savings in the future in a form of a probability distribution from which a confidence level can be computed. Accordingly, in recent years, many researchers have focused on these calibration approaches to enhance their performance in terms of precision and computational efficiency. In this thesis, a whole calibration methodology was studied and a thorough literature review is presented in chapter 1. In the subsequent chapters, each topic is analysed and elaborated in more details.

Sensitivity analysis is one of the main steps in the calibration methodology. It is conducted to select the most influential parameters on which calibration should be performed. In chapter 2, two methods from literature (Morris and RBD-FAST) are tested due to their computational efficiency and accuracy. These two methods are compared using Sobol method as the reference. The criteria selected for comparison are the methods precision in ranking the parameters accurately as ranked by Sobol method, their robustness in ranking all the parameters, the ones responsible for 95 % of the total variance, and the ones responsible for 90 % of the total variance. Moreover, their computational efficiency is also accounted for by assessing their performance with an increasing number of model evaluations. The indicators are mainly, the Pearson correlations coefficient and the Kendall tau coefficient. The case study for this analysis

is an individual house with 113 model parameters. The number of model evaluations needed by Sobol's method to rank the parameters is 460,000 corresponding to a sample size of 4000. RBD-FAST is found to rank with sufficient accuracy the parameters responsible of 90 % of the total variance with only hundreds of model evaluations. Morris' method performs even better in clustering these parameters. The difference is that with Morris, even the less important parameters are ranked with sufficient accuracy as Sobol method with only hundreds of model evaluations, which is not the case with RBD-FAST. All in all, Morris' method is found to be more robust and accurate than RBD-FAST, and it can be used with less risk; especially, if it needs to be followed with an identifiability analysis.

In chapter 3, different calibration methods (likelihood dependant and independent) are retained from literature and explained in details. The selected methods are applied on a virtual case study where the temperature profile is the data on which calibration is performed. The methods are compared in terms of accuracy and computational efficiency. The criteria used to undergo this comparison are the weighted Euclidean distance between the parameters' posteriors and their true values, and the average RMSE of the posteriors propagation. The results showed that APMC and Adams outperformed the other three algorithms. The reason behind this different performance is related to the sampler adaptation and not to the dependence on the likelihood function. ABC-RF showed a better performance with a lower number of model evaluations but the worst performance with an increasing number of simulations simulations when using the default hyper-parameters values for all the runs.

Another issue treated in this thesis is the identifiability of the calibration parameters. If the degree of interaction between the selected parameters is high, it will be harder for the calibration algorithm to converge. Accordingly, it is important to select not only the influential parameters but also the ones that have the least interactions. The identifiability concept is detailed in chapter 1. In chapter 4, a sensitivity-based identifiability analysis called orthogonalisation method is selected from literature and applied to a virtual case study. The aim is to study the effect of identifiability analysis through the assessment of the corresponding calibration performance compared to calibrating only the most influential parameters. The criteria are the identifiability of the parameters after calibration and the model predictive performance of the calibrated model. Accordingly, the first indicator is the Janson-Shannon distance computed between the priors and the posteriors to quantify the identifiability of the parameters separately. This indicator is only used under the conditions listed in chapter 4. The second indicator is the proposed distance measure ID, which computes the total identifiability of the model. The third

indicator is the deviance information criterion, which is used to evaluate the model predictive performance of the calibrated model. For the considered case study, it is shown that ranking the parameters using the orthogonalisation method is more appropriate than ranking them based on a sensitivity analysis only if few parameters are to be calibrated. Significant interaction may exist between the most important ones, which can be identified and accounted for through the identifiability analysis. If more parameters are included, both methods yield similar results.

This highlights the importance of an appropriate decision regarding the number of parameters to calibrate. This issue is also treated in chapter 4. The Janson Shannon and DIC are also used as indicators in this analysis. The aim of this study analysis is to assess the behaviour of calibration in terms of model predictive performance on the one hand and identifiability of the parameters on the other hand starting from calibrating only one parameter to calibrating 15 parameters. For the considered case study, it is found that calibrating the first three parameters is the most accurate, however, the calibration of the first eight parameters yielded almost similar precision, even though the first six parameters are found to be more identifiable than the rest.

The behaviour of ABC-RF found in chapter 3 triggered a deeper investigation on this method. In chapter 5 a new method called adaptive random forest (ARF) that is based on ABC-RF is proposed. This method integrates the sequential sampling used in other methods in ABC-RF. This method is applied to a virtual case study with five different priors: wide priors, precise priors, and priors shifted away from the true values. The results showed that this method can achieve very precise estimates for the true values and consequently a good model predictive performance with a small number of model evaluations (no more than 3000). The method was also compared to the ones presented in chapter 3 and it showed a considerably better computational efficiency.

Finally, all the methods were applied to a real case study with a real monitored temperature profile. The calibration yielded a better model predictive performance on both training and testing data with all the algorithms used in this chapter. ARF was found to be less accurate with an increasing number of iterations, which is not the case with the other algorithms. But again in this evaluation, the number of trees and leaves was not increased according to the number of simulations. More investigations need to be performed on ARF to enhance its performance in uncontrolled conditions, and to check its convergence. At the moment, without such a verification, APMC could be the best choice among the ones illustrated in this thesis. ARF has

shown potential compared to other methods with a limited number of model evaluations, but it has to be confirmed using different case studies.

This work leads to propose several perspectives in the future. The sensitivity methods used in this thesis are compared on one case study. It is obviously important to extend this study and do the same comparison on many different case studies in order to be able to draw more general conclusions. This is true for all the studies performed in this thesis, like the comparison of calibration methods, the application of the identifiability analysis, and the analysis regarding the number of parameters.

Away from extending the work of this thesis and replicate it on new case studies, it is also important to be aware of the limits in the methodologies. The comparison between the calibration methods was held by setting default values for the algorithms hyper-parameters, which are recommended by the authors. It is interesting to elaborate more on this and to tune the parameters of each algorithm and then do the comparison with the tuned hyper-parameters. This might be burdensome if the original model is used. Therefore, a metamodel is an option to bring down the computational cost. In this case, including the un-parallelisable NUTS sampler will pose no computational problems. Therefore, to include it in the comparison, the number of likelihood computations in each algorithms should be tracked, then to account for the computational efficiency, the time expected for each algorithm after considering parallelising could be estimated.

Another interesting aspect of calibration, which was not analysed in this thesis, is the data on which calibration is performed. In this thesis, only temperature profiles were used. In literature, some studies calibrate the model on temperature profiles, and some use the heat consumption as the basis for calibration. However, it seems like there is not clear analysis on what is more appropriate to do if both data are available. Moreover, if the model was calibrated on the heat consumption data, one wonders if it would perform well when used to predict the temperature profile and vice versa. This question is of importance e.g. when the model is calibrated based on the temperature profile in the summer where there is no heating power, and then used to predict the energy consumption in winter.

Moreover, the quantity and quality of the data on which calibration is performed is also a very important aspect of calibration. In many cases, the available data is uncertain or insufficient especially if occupancy-friendly sensors used. Accordingly, more research should

be performed to increase the reliability of the experimental campaigns on the one hand and to analyse the effect of the data quality and quantity on the whole inverse problem on the other hand.

ARF when applied to virtual data showed better performance than when applied to real measurements. It is one of the perspectives to investigate more on this method especially that it works very well in controlled conditions. Further work would be needed, particularly checking its statistical convergence e.g. using a statistical toy model, and increasing the number of trees and size of leaves according to the number of simulations.

References

Adams, M.P., Koh, E.J.Y., Vilas, M.P, Collier, C.J., Lambert, V.M., Sisson, S.A., Quiroz, M., Eve McDonald-Madden, McKenzie, L.J., and O'Brien, K.R. 2020. 'Predicting Seagrass Decline Due to Cumulative Stressors'. *Environmental Modelling & Software* 130 (August): 104717. https://doi.org/10.1016/j.envsoft.2020.104717.

Ahmad M., and Culp C.H. 2006. 'Uncalibrated Building Energy Simulation Modeling Results'. *HVAC&R Research* 12 (4): 1141–55. https://doi.org/10.1080/10789669.2006.10391455.

Akaike, H. 1974. 'A New Look at the Statistical Model Identification'. *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Akkari, S., Schalbart, P., and Peuportier, B. 2022. 'Assessment of Multiple Advanced Bayesian Calibration Algorithms in Building Energy Models'. *Conference IBPSA France*, May 2022.

Mara, T.A., and Joseph, O.R. 2008. 'Comparison of Some Efficient Methods to Evaluate the Main Effect of Computer Model Factors'. *Journal of Statistical Computation and Simulation* 78 (2): 167–78. https://doi.org/10.1080/10629360600964454.

Amit, Y., and Geman, D. 1997. 'Shape Quantization And Recognition With Randomized Trees.' *Neural Computation* 9 (October): 1545–88. https://doi.org/10.1162/neco.1997.9.7.1545.

Beaumont, M.A. 2010. 'Approximate Bayesian Computation in Evolution and Ecology'. *Annual Review of Ecology, Evolution, and Systematics* 41 (1): 379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beaumont, M.A., Cornuet, J.M., Marin, J.M, and Robert, CP. 2009. 'Adaptive Approximate Bayesian Computation'. *Biometrika* 96 (4): 983–90. https://doi.org/10.1093/biomet/asp052.

Beaumont, M.A, Zhang, W., and Balding, D.J. 2002. 'Approximate Bayesian Computation in Population Genetics'. *Genetics* 162 (4): 2025–35. https://doi.org/10.1093/genetics/162.4.2025.

Bellu, G., Saccomani, M., Audoly, S., and D'Angiò, L. 2007. 'DAISY: A New Software Tool to Test Global Identifiability of Biological and Physiological Systems'. *Computer Methods and Programs in Biomedicine* 88 (November): 52–61. https://doi.org/10.1016/j.cmpb.2007.07.002.

Biau, G., and Scornet, E. 2015. 'A Random Forest Guided Tour'. *ArXiv:1511.05741 [Math, Stat]*, November. http://arxiv.org/abs/1511.05741.

Blum, M.G.B., and François, O. 2010. 'Non-Linear Regression Models for Approximate Bayesian Computation'. *Statistics and Computing* 20 (1): 63–73. https://doi.org/10.1007/s11222-009-9116-0.

Booth, A.T., Choudhary, R., and Spiegelhalter, D.J. 2012. 'Handling Uncertainty in Housing Stock Models'. *Building and Environment* 48 (February): 35–47. https://doi.org/10.1016/j.buildenv.2011.08.016.

Bortot, P., Coles S.G., and Sisson, S.A., 2007. 'Inference for Stereological Extremes'. *Journal of the American Statistical Association* 102 (477): 84–92. https://doi.org/10.1198/016214506000000988.

Breiman, L., Friedman, J., Olshen, R., and Stone, C.J. 1983. 'Classification and Regression Trees'. https://doi.org/10.2307/2530946.

Breiman, L. 1996. 'Bagging Predictors'. *Machine Learning* 24 (2): 123–40. https://doi.org/10.1007/BF00058655.

Breiman, L. 2001. 'Random Forests'. *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brun, A., Clara S., and Wurtz, E. 2009. 'Analyse Du Comportement De Differents Codes De Calcul Dans Le Cas De Batiments À Haute Efficacite Energetique'. 9th colloque interuniversitaire Franco-québécois, Lille, 18-19 May 2009.

Brun, R., Reichert, P., and Künsch, H.R. 2001. 'Practical Identifiability Analysis of Large Environmental Simulation Models'. *Water Resources Research* 37 (4): 1015–30. https://doi.org/10.1029/2000WR900350.

Ching, J., and Chen, Y. 2007. 'Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging'. *Journal of Engineering Mechanics* 133 (7): 816–32. https://doi.org/10.1061/(ASCE)0733-9399(2007)133:7(816).

Chong, A., Lam, K.P., Pozzi, M., and Yang, J. 2017. 'Bayesian Calibration of Building Energy Models with Large Datasets'. *Energy and Buildings* 154 (November): 343–55. https://doi.org/10.1016/j.enbuild.2017.08.069.

Chong, A., and Menberg, K. 2018. 'Guidelines for the Bayesian Calibration of Building Energy Models'. *Energy and Buildings* 174 (September): 527–47. https://doi.org/10.1016/j.enbuild.2018.06.028.

Clarke, J.A., Yaneski, P.P., Pinney, A.A. 1991. *Harmonization of Thermal Properties of Building Materials*. Building Environmental Performance Analysis Club.

Coakley, D., Raftery, P., and Keane, M. 2014. 'A Review of Methods to Match Building Energy Simulation Models to Measured Data'. *Renewable and Sustainable Energy Reviews* 37 (September): 123–41. https://doi.org/10.1016/j.rser.2014.05.007.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., and Schaibly, J.H., 1973. 'A Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I. Theory': Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/AD0762420.

Wit, M.S. 2001. 'Uncertainty in Predictions of Thermal Comfort in Buildings'. https://repository.tudelft.nl/islandora/object/uuid%3Aa231bca8-ec81-4e22-8b34-4bafc062950e.

Diamond, S.C., Cappiello, C.C., and Hunn, B.D. 1986. 'DOE-2 Verification Project. Phase I. Final Report'. LA-10649-MS. Los Alamos National Lab., NM (USA). https://www.osti.gov/biblio/6025484.

Domínguez-Muñoz, F., Anderson, B., Cejudo-López, J.M, and Carrillo-Andrés, A. 2010. 'Uncertainty in the Thermal Conductivity of Insulation Materials'. *Energy and Buildings* 42 (11): 2159–68. https://doi.org/10.1016/j.enbuild.2010.07.006.

Domínguez-Muñoz, F., Cejudo-López, J.M., and Carrillo-Andrés, A. 2010. 'Uncertainty in Peak Cooling Load Calculations'. *Energy and Buildings* 42 (7): 1010–18. https://doi.org/10.1016/j.enbuild.2010.01.013.

Driscoll, E.A., and Landrum, D.B. 2004. 'Uncertainty Analysis on Heat Transfer Correlations for RP-1 Fuel in Copper Tubing'. In . Las Vegas, NV. https://ntrs.nasa.gov/citations/20040076962.

Drovandi, C.C., and Pettitt, A.N. 2011. 'Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation'. *Biometrics* 67 (1): 225–33. https://doi.org/10.1111/j.1541-0420.2010.01410.x.

Efron, B. 1979. 'Bootstrap Methods: Another Look at the Jackknife'. *The Annals of Statistics* 7 (1): 1–26.

Fan, H.H., and Kubatko, L.S. 2011. 'Estimating Species Trees Using Approximate Bayesian Computation'. *Molecular Phylogenetics and Evolution* 59 (2): 354–63. https://doi.org/10.1016/j.ympev.2011.02.019.

Filippi S., Barnes, C.P., Cornebise, J., and Stumpf, M.P.H., 2013. 'On Optimality of Kernels for Approximate Bayesian Computation Using Sequential Monte Carlo'. *Statistical Applications in Genetics and Molecular Biology* 12 (1). https://doi.org/10.1515/sagmb-2012-0069.

Gábor, A., Villaverde, A.F., and Banga, J.R. 2017. 'Parameter Identifiability Analysis and Visualization in Large-Scale Kinetic Models of Biosystems'. *BMC Systems Biology* 11 (1): 54. https://doi.org/10.1186/s12918-017-0428-y.

Gatelli, D., Kucherenko, S., Ratto, M., and Tarantola, S. 2009. 'Calculating First-Order Sensitivity Measures: A Benchmark of Some Recent Methodologies'. *Reliability Engineering & System Safety* 94 (7): 1212–19. https://doi.org/10.1016/j.ress.2008.03.028.

Gelman, A., Hwang, J., and Vehtari, A. 2013. 'Understanding Predictive Information Criteria for Bayesian Models'. *ArXiv:1307.5928 [Stat]*, July. http://arxiv.org/abs/1307.5928.

Genuer, R., Poggi, J., and Tuleau, C. 2008. 'Random Forests: Some Methodological Insights'. *ArXiv:0811.3619 [Stat]*, November. http://arxiv.org/abs/0811.3619.

Genuer, R., Poggi, J., and Tuleau-Malot, C. 2010. 'Variable Selection Using Random Forests'. *Pattern Recognition Letters* 31 (14): 2225–36. https://doi.org/10.1016/j.patrec.2010.03.014.

Goffart, J. 2013. 'Impact de la variabilité des données météorologiques sur une maison basse consommation. Application des analyses de sensibilité pour les entrées temporelles.' PhD thesis, Université de Grenoble. https://tel.archives-ouvertes.fr/tel-00982150.

Goffart, J., Rabouille, M., and Mendes, N. 2017. 'Uncertainty and Sensitivity Analysis Applied to Hygrothermal Simulation of a Brick Building in a Hot and Humid Climate'. *Journal of*

Building Performance Simulation 10 (1): 37–57. https://doi.org/10.1080/19401493.2015.1112430.

Goffart, J., and Woloszyn, M. 2021. 'EASI RBD-FAST: An Efficient Method of Global Sensitivity Analysis for Present and Future Challenges in Building Performance Simulation'. *Journal of Building Engineering* 43 (November): 103129. https://doi.org/10.1016/j.jobe.2021.103129.

Harrison, J.U., and Baker, R.E. 2020. 'An Automatic Adaptive Method to Combine Summary Statistics in Approximate Bayesian Computation'. *PLOS ONE* 15 (8): e0236954. https://doi.org/10.1371/journal.pone.0236954.

Heo, Y., Choudhary, R., and Augenbroe. G.A. 2012. 'Calibration of Building Energy Models for Retrofit Analysis under Uncertainty'. *Energy and Buildings* 47 (April): 550–60. https://doi.org/10.1016/j.enbuild.2011.12.029.

Heo, Y., Graziano, D.J., Guzowski, L., and Muehleisen, R.T., 2015. 'Evaluation of Calibration Efficacy under Different Levels of Uncertainty'. *Journal of Building Performance Simulation* 8 (3): 135–44. https://doi.org/10.1080/19401493.2014.896947.

Hickerson, M.J., Stahl, E.A., and Lessios, H.A. 2006. 'Test for Simultaneous Divergence Using Approximate Bayesian Computation'. *Evolution* 60 (12): 2435–53. https://doi.org/10.1111/j.0014-3820.2006.tb01880.x.

Ho, T.K. 1995. 'Random Decision Forests'. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.

Ho, T.K. 1998. 'The Random Subspace Method for Constructing Decision Forests'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832–44. https://doi.org/10.1109/34.709601.

Hoffman, M.D., and Gelman, A. 2011. 'The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo'. *ArXiv:1111.4246 [Cs, Stat]*, November. http://arxiv.org/abs/1111.4246.

Homma, T., and Saltelli, A. 1996. 'Importance Measures in Global Sensitivity Analysis of Nonlinear Models'. *Reliability Engineering & System Safety* 52 (1): 1–17. https://doi.org/10.1016/0951-8320(96)00002-6.

Jansen, M.J.W. 1999. 'Analysis of Variance Designs for Model Output'. *Computer Physics Communications* 117 (1–2): 35–43. https://doi.org/10.1016/S0010-4655(98)00154-4.

Johnston, S.T., Simpson, M.J., McElwain, D.L.S, Binder, B.J, and Ross, J.V. 2014. 'Interpreting Scratch Assays Using Pair Density Dynamics and Approximate Bayesian Computation'. *Open Biology* 4 (9): 140097. https://doi.org/10.1098/rsob.140097.

Juricic, S. 2020. 'Identifiability of the Thermal Performance of a Building Envelope from Poorly Informative Data'. PhD thesis, Thermics [physics.class-ph]. Université Savoie Mont Blanc, 2020. English. NNT: 2020CHAMA014. tel-03181809.

Kang, Y., and Krarti, M. 2016. ‘Bayesian-Emulator Based Parameter Identification for Calibrating Energy Models for Existing Buildings’. *Building Simulation* 9 (4): 411–28. https://doi.org/10.1007/s12273-016-0291-6.

Kennedy, M.C., and O’Hagan, A. 2001. ‘Bayesian Calibration of Computer Models’. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3): 425–64. https://doi.org/10.1111/1467-9868.00294.

Kim, Y., and Park, C. 2016. ‘Stepwise Deterministic and Stochastic Calibration of an Energy Simulation Model for an Existing Building’. *Energy and Buildings* 133 (December): 455–68. https://doi.org/10.1016/j.enbuild.2016.10.009.

Kristensen, M.H., Choudhary R., Pedersen, R.H., and Petersen, S. 2017. ‘Bayesian Calibration Of Residential Building Clusters Using A Single Geometric Building Representation’. 5th IBPSA Building Simulation Conference, San Francisco, August.

Lavigne, K. 2014. ‘Assisted Calibration in Building Simulation–Algorithm Description and Case Studies’. Text. AIVC. 23 June 2014. https://www.aivc.org/resource/assisted-calibration-building-simulation-algorithm-description-and-case-studies.

Lavigne, K. 2009. ‘Assisted Calibration in Building Simulation–Algorithm Description and Case Studies’. Building Simulation 2009 conference (IBPSA), Glascow, July.

Lee, P., Lam, P.T.I, Yik, F.W.H, and Chan, E.H.W. 2013. ‘Probabilistic Risk Assessment of the Energy Saving Shortfall in Energy Performance Contracting Projects–A Case Study’. *Energy and Buildings* 66 (November): 353–63. https://doi.org/10.1016/j.enbuild.2013.07.018.

Lenormand, M., Jabot, F., and Deffuant, G. 2013. ‘Adaptive Approximate Bayesian Computation for Complex Models’. *Computational Statistics* 28 (6): 2777–96. https://doi.org/10.1007/s00180-013-0428-3.

Lim, H., and Zhai, Z.J. 2017. ‘Comprehensive Evaluation of the Influence of Meta-Models on Bayesian Calibration’. *Energy and Buildings* 155 (November): 66–75. https://doi.org/10.1016/j.enbuild.2017.09.009.

Lintusaari, J., Gutmann, M.U., Dutta, R., Kaski, S., and Corander, J. 2016. ‘Fundamentals and Recent Developments in Approximate Bayesian Computation’. *Systematic Biology*, October, syw077. https://doi.org/10.1093/sysbio/syw077.

Macdonald, I.A. 2002. ‘Quantifying the effects of uncertainty in building simulation’. PhD thesis, Glasgow, Scotland: University of Strathclyde. Dept. of Mechanical Engineering.https://www.strath.ac.uk/media/departments/mechanicalengineering/esru/research/phdmphilprojects/macdonald_thesis.pdf.

Mara, T.A. 2009. ‘Extension of the RBD-FAST Method to the Computation of Global Sensitivity Indices’. *Reliability Engineering & System Safety* 94 (8): 1274–81. https://doi.org/10.1016/j.ress.2009.01.012.

Marchi, B., and Zanoni, S. 2017. ‘Supply Chain Management for Improved Energy Efficiency: Review and Opportunities’. *Energies* 10 (10): 1618. https://doi.org/10.3390/en10101618.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. 2003. 'Markov Chain Monte Carlo without Likelihoods'. *Proceedings of the National Academy of Sciences* 100 (26): 15324–28. https://doi.org/10.1073/pnas.0306899100.

Mechri, H.E., Capozzoli, A., and Corrado, V. 2010. 'USE of the ANOVA Approach for Sensitive Building Energy Design'. *Applied Energy* 87 (10): 3073–83. https://doi.org/10.1016/j.apenergy.2010.04.001.

Menberg, K., Heo, Y., and Choudhary. R. 2017. 'Efficiency and Reliability of Bayesian Calibration of Energy Supply System Models', 10.

Miao, H., Xia, X., Perelson, A.S., and Wu, H. 2011. 'On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics'. *SIAM Review* 53 (1): 3–39. https://doi.org/10.1137/090757009.

Minson, S.E., Simons, M., and Beck, J.L. 2013. 'Bayesian Inversion for Finite Fault Earthquake Source Models I—Theory and Algorithm'. *Geophysical Journal International* 194 (3): 1701–26. https://doi.org/10.1093/gji/ggt180.

Moon, H. 2009. 'Assessing Mold Risks in Buildings under Uncertainty'. *Undefined*. https://www.semanticscholar.org/paper/Assessing-Mold-Risks-in-Buildings-under-Uncertainty-Moon/c63d8df34f290e6f41774b4207a428645cc60fcc.

Morris, M.D. 1991. 'Factorial Sampling Plans for Preliminary Computational Experiments'. *Technometrics* 33 (2): 161–74. https://doi.org/10.1080/00401706.1991.10484804.

Muehleisen, R.T., and Bergerson, J. 2016. 'Bayesian Calibration - What, Why And How', International High Performance Buildings Conference. Paper 167. Purdue. http://docs.lib.purdue.edu/ihpbc/167.

Munaretto, F. 2014. 'Étude de l'influence de l'inertie thermique sur les performances énergétiques des bâtiments'. PhD., Ecole Nationale Supérieure des Mines de Paris. https://pastel.archives-ouvertes.fr/pastel-01068784.

O'Neill, Z., Eisenhower, B., Yuan, S., Bailey, T., Narayanan, S., and Fonoberov, V. 2011. 'Modeling and Calibration of Energy Models for a DoD Building'. In . Vol. 117. *ASHRAE journal*.

Pannier, M. 2017. 'Etude de La Quantification Des Incertitudes En Analyse de Cycle de Vie Des Bâtiments'. PhD thesis, MINES ParisTech PSL.

Pannier, M, Schalbart, P., and Peuportier, B. 2018. 'Comprehensive Assessment of Sensitivity Analysis Methods for the Identification of Influential Factors in Building Life Cycle Assessment'. *Journal of Cleaner Production* 199 (October): 466–80. https://doi.org/10.1016/j.jclepro.2018.07.070.

Peuportier, B. 2005. 'Bancs d'essais de Logiciels de Simulation Thermique.', Journée thématique IBPSA France, La Rochelle.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. 1999. 'Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites'. *Molecular*

Biology and Evolution 16 (12): 1791–98. https://doi.org/10.1093/oxfordjournals.molbev.a026091.

Pudlo, P., Marin, J., Estoup, A., Cornuet, J., Gautier, M., and Robert, C.P. 2015. 'Reliable ABC Model Choice via Random Forests'. *ArXiv:1406.6288 [q-Bio, Stat]*, September. http://arxiv.org/abs/1406.6288.

Quaiser, T., and Mönnigmann, M. 2009. 'Systematic Identifiability Testing for Unambiguous Mechanistic Modeling – Application to JAK-STAT, MAP Kinase, and NF-κ B Signaling Pathway Models'. *BMC Systems Biology* 3 (1): 50. https://doi.org/10.1186/1752-0509-3-50.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. 2009. 'Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood'. *Bioinformatics* 25 (15): 1923–29. https://doi.org/10.1093/bioinformatics/btp358.

Raynal, L., Marin, J., Pudlo, P., Ribatet, M., Robert, C.P., and Estoup, A. 2017. 'ABC Random Forests for Bayesian Parameter Inference'. *Peer Community in Evolutionary Biology*, November, 100036. https://doi.org/10.24072/pci.evolbiol.100036.

Reddy, A. 2006. 'Literature Review on Calibration of Building Energy Simulation Programs: Uses, Problems, Procedure, Uncertainty, and Tools'. *ASHRAE Transactions* 112 (January): 226–40.

Robillart, M. 2015. 'Etude de stratégies de gestion en temps réel pour des bâtiments énergétiquement performants'. PhD thesis, Ecole Nationale Supérieure des Mines de Paris. https://pastel.archives-ouvertes.fr/tel-01299525.

Rouchier, S. 2018. 'Solving Inverse Problems in Building Physics: An Overview of Guidelines for a Careful and Optimal Use of Data'. *Energy and Buildings* 166 (February). https://doi.org/10.1016/j.enbuild.2018.02.009.

Royapoor, M., and Roskilly, T. 2015. 'Building Model Calibration Using Energy and Environmental Data'. *Energy and Buildings* 94 (May): 109–20. https://doi.org/10.1016/j.enbuild.2015.02.050.

Ruggeri, A.G., Gabrielli, L., and Scarpa, M. 2020. 'Energy Retrofit in European Building Portfolios: A Review of Five Key Aspects'. *Sustainability* 12 (18): 7465. https://doi.org/10.3390/su12187465.

Saltelli, A., ed. 2008. *Global Sensitivity Analysis: The Primer*. Chichester, England ; Hoboken, NJ: John Wiley.

Saltelli, A., Tarantola, S., and Chan, K.P.S. 1999. 'A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output'. *Technometrics* 41 (1): 39–56. https://doi.org/10.1080/00401706.1999.10485594.

Saltelli, A., ed. 2008. *Sensitivity Analysis*. Paperback ed. Wiley Paperback Series. Chichester: Wiley.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. 2010. 'Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity

Index'. *Computer Physics Communications* 181 (2): 259–70. https://doi.org/10.1016/j.cpc.2009.09.018.

Saltelli, A., and Bolado, R. 1998. 'An Alternative Way to Compute Fourier Amplitude Sensitivity Test (FAST)'. *Computational Statistics & Data Analysis* 26 (4): 445–60. https://doi.org/10.1016/S0167-9473(97)00043-1.

Sandberg, N.H., Sartori, I., Heidrich, O., Dawson, R., Dascalaki, E., Dimitriou, S., Vimm-r, T., et al. 2016. 'Dynamic Building Stock Modelling: Application to 11 European Countries to Support the Energy Efficiency and Retrofit Ambitions of the EU'. *Energy and Buildings* 132 (November): 26–38. https://doi.org/10.1016/j.enbuild.2016.05.100.

Satterthwaite, F.E. 1959. 'Random Balance Experimentation'. *Technometrics* 1 (2): 111–37. https://doi.org/10.2307/1266466.

Shen, H., and Tzempelikos, A. 2013. 'Sensitivity Analysis on Daylighting and Energy Performance of Perimeter Offices with Automated Shading'. *Building and Environment* 59 (January): 303–14. https://doi.org/10.1016/j.buildenv.2012.08.028.

Silva, A.S., and Ghisi, E. 2014. 'Uncertainty Analysis of User Behaviour and Physical Parameters in Residential Building Performance Simulation'. *Energy and Buildings* 76 (June): 381–91. https://doi.org/10.1016/j.enbuild.2014.03.001.

Sisson, S.A., Fan, Y. and Beaumont, M.A. 2018. 'Overview of Approximate Bayesian Computation'. *ArXiv:1802.09720 [Stat]*, February. http://arxiv.org/abs/1802.09720.

Sisson, S.A., Fan Y., and Tanaka, M.M. 2007. 'Sequential Monte Carlo without Likelihoods'. *Proceedings of the National Academy of Sciences* 104 (6): 1760–65. https://doi.org/10.1073/pnas.0607208104.

Sobol, I.M., and Shukman, B.V. 1993. 'Random and Quasirandom Sequences: Numerical Estimates of Uniformity of Distribution'. *Mathematical and Computer Modelling* 18 (8): 39–45. https://doi.org/10.1016/0895-7177(93)90160-Z.

Sobol, I.M., Tarantola S., Gatelli, D., Kucherenko, S.S., and Mauntz, W. 2007. 'Estimating the Approximation Error When Fixing Unessential Factors in Global Sensitivity Analysis'. *Reliability Engineering & System Safety* 92 (7): 957–60. https://doi.org/10.1016/j.ress.2006.07.001.

Sokol, J., Davila, C.C., and Reinhart, C.F. 2017. 'Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models'. *Energy and Buildings* 134 (January): 11–24. https://doi.org/10.1016/j.enbuild.2016.10.050.

Spitz, C. 2012. 'Analyse de la fiabilité des outils de simulation et des incertitudes de métrologie appliquée à l'efficacité énergétique des bâtiments'. PhD thesis, Université de Grenoble. https://tel.archives-ouvertes.fr/tel-00768506.

Spitz, C, Mora, L., Wurtz, E., and Jay, A. 2012. 'Practical Application of Uncertainty Analysis and Sensitivity Analysis on an Experimental House'. *Energy and Buildings* 55 (December): 459–70. https://doi.org/10.1016/j.enbuild.2012.08.013.

Sun, Y. 2014. ‘Closing the Building Energy Performance Gap by Improving Our Predictions’, June. https://smartech.gatech.edu/handle/1853/52285.

Sun, Y., Heo, Y., Tan, M., Xie, H., Jeff Wu, C.F., and Augenbroe, G. 2014. ‘Uncertainty Quantification of Microclimate Variables in Building Energy Models’. *Journal of Building Performance Simulation* 7 (1): 17–32. https://doi.org/10.1080/19401493.2012.757368.

Tarantola, S., Gatelli, D., and Mara, T.A. 2006. ‘Random Balance Designs for the Estimation of First Order Global Sensitivity Indices’. *Reliability Engineering & System Safety* 91 (6): 717–27. https://doi.org/10.1016/j.ress.2005.06.003.

Thevenard, D., and Haddad, K. 2006. ‘Ground Reflectivity in the Context of Building Energy Simulation’. *Energy and Buildings* 38 (8): 972–80. https://doi.org/10.1016/j.enbuild.2005.11.007.

Tian, W. 2013. ‘A Review of Sensitivity Analysis Methods in Building Energy Analysis’. *Renewable and Sustainable Energy Reviews* 20 (April): 411–19. https://doi.org/10.1016/j.rser.2012.12.014.

Tian, W., Heo Y., De Wilde, P., Li, Z., Yan, D., Park, C.S., Feng, X., and Augenbroe, G. 2018. ‘A Review of Uncertainty Analysis in Building Energy Assessment’. *Renewable and Sustainable Energy Reviews* 93 (October): 285–301. https://doi.org/10.1016/j.rser.2018.05.029.

Tissot, J., and Prieur, C. 2012. ‘Bias Correction for the Estimation of Sensitivity Indices Based on Random Balance Designs’. *Reliability Engineering & System Safety* 107 (November): 205–13. https://doi.org/10.1016/j.ress.2012.06.010.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M.P.H. 2009. ‘Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems’. *Journal of The Royal Society Interface* 6 (31): 187–202. https://doi.org/10.1098/rsif.2008.0172.

Tuominen, Pekka, Krzysztof Klobut, Anne Tolman, Afi Adjei, and Marjolein de Best-Waldhober. 2012. ‘Energy Savings Potential in Buildings and Overcoming Market Barriers in Member States of the European Union’. *Energy and Buildings* 51 (August): 48–55. https://doi.org/10.1016/j.enbuild.2012.04.015.

Turner, B.M., and Van Zandt, T. 2012. ‘A Tutorial on Approximate Bayesian Computation’. *Journal of Mathematical Psychology* 56 (2): 69–85. https://doi.org/10.1016/j.jmp.2012.02.005.

Vorger, E., Schalbart, P., and Peuportier, B. 2014. ‘Integration of a Comprehensive Stochastic Model of Occupancy in Building Simulation to Study How Inhabitants Influence Energy Performance’. In *30th International Plea 2014 Conference*. Ahmedabad, India. https://hal-mines-paristech.archives-ouvertes.fr/hal-01460068.

Wang, Q., Augenbroe, G., and Sun, Y. 2014. ‘The Role of Construction Detailing and Workmanship in Achieving Energy-Efficient Buildings’. In *Construction Research Congress 2014*, 2224–33. Atlanta, Georgia: American Society of Civil Engineers. https://doi.org/10.1061/9780784413517.226.

Wilkinson, R. 2013. ‘Approximate Bayesian Computation (ABC) Gives Exact Results under the Assumption of Model Error’. *Statistical Applications in Genetics and Molecular Biology* 12 (May): 1–13. https://doi.org/10.1515/sagmb-2013-0010.

Xu, C., and Gertner, G.Z. 2008. ‘A General First-Order Global Sensitivity Analysis Method’. *Reliability Engineering & System Safety* 93 (7): 1060–71. https://doi.org/10.1016/j.ress.2007.04.001.

Yan, W., and Goebel, K. 2004. ‘Designing Classifier Ensembles with Constrained Performance Requirements’. *Proceedings of SPIE - The International Society for Optical Engineering* 5434 (April). https://doi.org/10.1117/12.542616.

Yao, K.Z., Shaw, B.M., Kou, B., McAuley, K.B., and Bacon, D.W. 2003. ‘Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design’. *Polymer Reaction Engineering* 11 (3): 563–88. https://doi.org/10.1081/PRE-120024426.

Yildiz, Y., Korkmaz, K., Göksal Özbalta, T., and Durmus Arsan, Z. 2012. ‘An Approach for Developing Sensitive Design Parameter Guidelines to Reduce the Energy Requirements of Low-Rise Apartment Buildings’. *Applied Energy* 93 (May): 337–47. https://doi.org/10.1016/j.apenergy.2011.12.048.

Yoon, S., Park, C., and Augenbroe, G. 2011. ‘On-Line Parameter Estimation and Optimal Control Strategy of a Double-Skin System’. *Building and Environment* 46 (5): 1141–50. https://doi.org/10.1016/j.buildenv.2010.12.001.

Zhang, Y., Van Bael, A., Andrade-Campos, A., and Coppieters, S. 2022. ‘Parameter Identifiability Analysis: Mitigating the Non-Uniqueness Issue in the Inverse Identification of an Anisotropic Yield Function’. *International Journal of Solids and Structures* 243 (May): 111543. https://doi.org/10.1016/j.ijsolstr.2022.111543.

Zhu, C., Tian, W., Yin, B., Li, Z., and Shi, J. 2020. ‘Uncertainty Calibration of Building Energy Models by Combining Approximate Bayesian Computation and Machine Learning Algorithms’. *Applied Energy* 268 (June): 115025. https://doi.org/10.1016/j.apenergy.2020.115025.

Appendices

Appendix A. Morris' method

Morris is a screening sensitivity method. With this method, each parameter is assigned a uniform distribution with a lower and upper bound, unlike variance based methods, where they can take advantage of the probability distributions assigned to each parameter. The parameters are then reduced to be dimensionless varying between 0 and 1.

The parameter space is then discretised based on a pre-identified p value that corresponds to the number of discretisation levels associated to each parameter θ_i with the step $\frac{1}{p-1}$. The discretised space for each parameter can then be expressed as a set as shown in equation (A.1). The set size is the number of levels p indicated. Only values from this set are drawn.

$$\theta_{set} = \left\{0, \frac{1}{p-1}, \frac{2}{p-2}, \dots, 1\right\}. \quad \text{(A.1)}$$

The methodology is to assign base values for all the parameters and to carry out one simulation run. Then, one parameter is perturbed keeping all the others at their base values and another model run is executed. At the next step, another parameter is perturbed keeping the previous parameter at its perturbed value and all the others at their base. The base values and the perturbations correspond to random draws from the set of each parameter space θ_{set}. This perturbation is done via a defined step that is proposed by Morris (1991) to be estimated via $\Delta = \frac{p}{2(p-1)}$ with p being an even number.

This is then done for all the parameters (in random order) until all of them are perturbed summing up to $(k+1)$ number of model runs where k is the parameters number. This is repeated r times. This leads to $r(k+1)$ number of model runs to carry out this method.

The influence of each parameter on the model output is estimated by calculating the change of the model output relative to the change of the model parameter. This is called the elementary effect EE_j of the parameter j. The elimentary effect EE_j^i for a parameter j in the repetition i can be calculated as follows:

$$EE_j^i = \frac{Y(\theta_1^i,\ \theta_2^i, \dots, \theta_j^i + \Delta, \dots \theta_k^i) - Y(\theta_1^i, \dots, \theta_j^i, \dots, \theta_k^i)}{\Delta\theta} \tag{A.2}$$

This effect is calculated for the same parameter j in each repetition i until r values of the elementary effects compose a sample for the parameter. Then the absolute mean and the variance of this sample are calculated:

$$\mu_j^* = \frac{1}{r}\sum_{i=1}^{r}\left|EE_j^i\right| \tag{A.3}$$

$$\sigma_j = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}\left(EE_j^i - \mu_j\right)^2} \tag{A.4}$$

where μ_j^* is the sample absolute mean, and μ_j is the sample mean.

Plotting the values of σ_j as a function of μ_j^* for all the parameters allows to classify the parameters into three different groups (Figure A.1):

- Parameters with negligible effects (shown as group 1)
- Influential parameters with linear effects and no interactions (shown in group 2)
- Influential parameters with nonlinearities and/or interaction effects (shown in group 3).

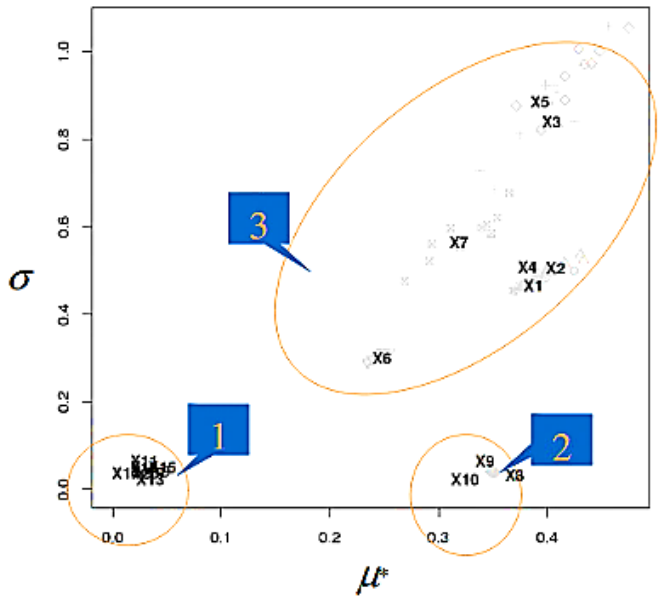

Figure A.1: Groups of parameters in Morris method (Iooss, 2011)

In order to obtain a ranking of the parameters according to their influence, considering the linear and additive effects and non-linear and interaction effects, the Euclidean distance to the origin of the graph for each parameter is considered (Recht et al., 2014). The further the point corresponding to a parameter is from the origin on the graph (i.e. the larger μ*, or the larger σ), the more influential the parameter is on the results.

$$d_j^* = \sqrt{{\mu_j^*}^2 + \sigma_i^2} \tag{A.5}$$

Appendix B. Sobol indices

There exist different estimators that evaluate the Sobol sensitivity indices. Here, the demarche of computing these indices is presented. Two matrices *A* and *B* each of dimension (*N*, *k*) are generated. The i^{th} row in both matrices corresponds to the same parameter but with different values. The values for each parameter are sampled from its distribution function. We can consider *A* as the sampling matrix and at *B* as the resampling matrix. For estimating the variance caused by a parameter θ_i a new matrix C_i is created of the same dimensions consisting of the same values of matrix *B* except for the i^{th} column (parameter under investigation) where it is taken from *A*. Thus, it can be stated that the new matrix C_i (annotated ${B_A}^{(i)}$) resamples all the parameters except the parameter that is to be studied. The multiplication of these two functions results in an intuitive justification whether the parameter under question is important or not. When x_i is uninfluential, low and high f values will be randomly multiplied by each other yielding a low D_i, thus a low S_i. On the contrary if θ_i is important, then low values of f will be multiplied by low values of the second function, and high values will be multiplied by high values resulting in a high D_i and accordingly a high S_i.

$$\mathrm{A} = \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_{\mathrm{i}}^{(1)} & \dots & \theta_{\mathrm{k}}^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_{\mathrm{i}}^{(2)} & \dots & \theta_{\mathrm{k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_1^{(\mathrm{N}-1)} & \theta_2^{(\mathrm{N}-1)} & \dots & \theta_{\mathrm{i}}^{(\mathrm{N}-1)} & \dots & \theta_{\mathrm{k}}^{(\mathrm{N}-1)} \\ \theta_1^{(\mathrm{N})} & \theta_2^{(\mathrm{N})} & \dots & \theta_{\mathrm{i}}^{(\mathrm{N})} & \dots & \theta_{\mathrm{k}}^{(\mathrm{N})} \end{bmatrix} \tag{B.1}$$

$$\mathrm{B} = \begin{bmatrix} \theta_{\mathrm{k}+1}^{(1)} & \theta_{\mathrm{k}+2}^{(1)} & \dots & \theta_{\mathrm{k}+\mathrm{i}}^{(1)} & \dots & \theta_{2\mathrm{k}}^{(1)} \\ \theta_{\mathrm{k}+1}^{(2)} & \theta_{\mathrm{k}+2}^{(2)} & \dots & \theta_{\mathrm{k}+\mathrm{i}}^{(2)} & \dots & \theta_{2\mathrm{k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_{\mathrm{k}+1}^{(\mathrm{N}-1)} & \theta_{\mathrm{k}+2}^{(\mathrm{N}-1)} & \dots & \theta_{\mathrm{k}+\mathrm{i}}^{(\mathrm{N}-1)} & \dots & \theta_{2\mathrm{k}}^{(\mathrm{N}-1)} \\ \theta_{\mathrm{k}+1}^{(\mathrm{N})} & \theta_{\mathrm{k}+2}^{(\mathrm{N})} & \dots & \theta_{\mathrm{k}+\mathrm{i}}^{(\mathrm{N})} & \dots & \theta_{2\mathrm{k}}^{(\mathrm{N})} \end{bmatrix} \tag{B.2}$$

$$\mathrm{C_i} = \begin{bmatrix} \theta_{\mathrm{k}+1}^{(1)} & \theta_{\mathrm{k}+2}^{(1)} & \dots & \theta_{\mathrm{i}}^{(1)} & \dots & \theta_{2\mathrm{k}}^{(1)} \\ \theta_{\mathrm{k}+1}^{(2)} & \theta_{\mathrm{k}+2}^{(2)} & \dots & \theta_{\mathrm{i}}^{(2)} & \dots & \theta_{2\mathrm{k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_{\mathrm{k}+1}^{(\mathrm{N}-1)} & \theta_{\mathrm{k}+2}^{(\mathrm{N}-1)} & \dots & \theta_{\mathrm{i}}^{(\mathrm{N}-1)} & \dots & \theta_{2\mathrm{k}}^{(\mathrm{N}-1)} \\ \theta_{\mathrm{k}+1}^{(\mathrm{N})} & \theta_{\mathrm{k}+2}^{(\mathrm{N})} & \dots & \theta_{\mathrm{i}}^{(\mathrm{N})} & \dots & \theta_{2\mathrm{k}}^{(\mathrm{N})} \end{bmatrix} \tag{B.3}$$

Theses matrices are then solved as shown in the following formulation given by (Andrea Saltelli 2008):

$$D_i = \frac{1}{n}\sum_{m=1}^{n} f\left(\theta^1_{(\sim i)m}, \theta^1_{im}\right) f\left(\theta^2_{(\sim i)m}, \theta^1_{im}\right) - f_0^2 \tag{B.4}$$

where *n* is the number of samples and θ_{im} means the sample number *m* of the i^{th} parameter. The annotation $(\sim \mathrm{i})m$ corresponds to the samples at number *m* (iteration *m*) for each parameter except the i^{th} one. The superscripts (*1*) and (*2*) indicate the usage of two sampling matrices *A* and *B*.

Simulating the model with all the input values in the matrices *A*, *B*, and C_i yields three different output vectors (A), $Y(B)$, $Y(C_i)$. Accordingly, equation (B.4) can be written as follows:

$$D_i = \frac{1}{N} \times \left(\sum_{j=1}^{N} \left(Y(A)\right)^{(j)} \times \left(Y(C_i)\right)^{(j)} \right) - f_0^2 \tag{B.5}$$

The total variance can be formulated as follows:

$$D = \frac{1}{N} \times \left(\sum_{j=1}^{N} Y(A)^{(j)^2} \right) - f_0^2 \tag{B.6}$$

The Monte Carlo estimation of the integral $D(\sim i)$ needed to evaluate the total index is:

$$\frac{1}{N} \times \left(\sum_{j=1}^{N} \left(Y(B)\right)^{(j)} \times \left(Y(C_i)\right)^{(j)} \right) - f_0^2 \tag{B.7}$$

Equations (B.5), (B.6), and (B.7) can be replaced in equations (2.4) and (2.7). Consequently, the first and the total effect indices for all the parameters can be estimated using the following formulas:

$$S_i = \frac{\frac{1}{N} \times \left(\sum_{j=1}^{N} \left(Y(A)\right)^{(j)} \times \left(Y(C_i)\right)^{(j)} \right) - f_0^2}{\frac{1}{N} \times \left(\sum_{j=1}^{N} Y(A)^{(j)^2} \right) - f_0^2} \tag{B.8}$$

$$TS_i = 1 - \frac{\frac{1}{N} \times \left(\sum_{j=1}^{N} \left(Y(B)\right)^{(j)} \times \left(Y(C_i)\right)^{(j)}\right) - f_0^2}{\frac{1}{N} \times \left(\sum_{j=1}^{N} Y(A)^{(j)^2}\right) - f_0^2} \quad \text{(B.9)}$$

The parameter is considered non-influential if its total effect index TS_i is low and it can be fixed as a constant value without affecting on the model output. Such a parameter should not be accounted for in the calibration analysis as it just increases the complexity of the algorithms used with no benefit. If the sum of all the first order indices for all the parameters is nearly 1, the model is considered to be additive with no interaction between the different parameters where the difference $(1 - \sum s_i)$ is an indicator of the presence of interaction in the model. In the same manner, by subtracting S_i from TS_i of a parameter, it can be indicated whether this parameter has interaction with the other parameters or not and by how much. Another property about the indices is that if the summation of all the total effect indices of all the parameters is near one, then the model is considered to have few interactions between its parameters. The reason behind this is that the interaction between two parameters for example is computed once during the calculation of the index of the first parameter and another time during the estimation of the other parameter, thus the interactions between the same parameters are computed more than once during the whole process.

This method needs to simulate *N* samples twice (once for matrix *A* and once for matrix *B*) and then to simulate matrix C_i consisting of *N* samples *K* times for each parameter. Thus, the number of simulations needed to estimate all the indices is $2N + KN = N(K + 2)$. This is computationally expensive especially with complex models that consume a considerable amount of time to finish simulation.

Different studies have been dedicated to the improvement of the estimators. Jansen (1999) proposed modified estimators based on the same matrices for ST_i as follows:

$$TS_i = \frac{\frac{1}{2N} \sum_{j=1}^{N} \left(\left(Y(A)\right)^{(j)} - Y(A_B^{(i)})^{(j)}\right)^2}{D} \quad \text{(B.10)}$$

where $Y_{A_B^{(i)}}$ is the simulation output of a matrix $A_B^{(i)}$ which contrary to C_i comprises the samples in matrix A except for the i^{th} factor which is taken from matrix B. Saltelli et al. (2010) proved that the high order indices estimator of Jansen (1999) is the best estimator compared to those of Sobol' et al. (2007) and Homma and Saltelli (1996). Consequently it is used in this thesis.

Appendix C. Regression post-processing

Beaumont et al. (2002) refined the model proposed by Pritchard et al. (1999). Briefly, in this refined approach, the tolerance (or minimum threshold) is not specified beforehand. However, it is specified after simulating *M* particles in a way that a certain proportion of these particles are accepted. The idea is to weight the accepted samples θ of the parameters according to the difference between the simulation output and the observed data $S(Z) - S(Y_i)$. These weights are used to train a weighted linear regression on the accepted samples θ. These samples are corrected using this linear regression. This adjusts the parameters posteriors in order to weaken the effect of the specified tolerance.

They have used the Epanechnikov kernel function presented as follows to weight the accepted samples:

$$k_\delta(t) = \begin{cases} c\delta^{-1}(1-(t/\delta)^2) & t \le \delta \\ 0 & t > \delta \end{cases} \quad \text{(C.1)}$$

where *c* is a normalising constant. The regression model that is fit to the accepted samples is given as follows:

$$\theta_i = \alpha + (s_i - s)^T\beta + \varepsilon_i, i = 1, \dots, m, \quad \text{(C.2)}$$

where α is an intercept, β is a vector of regression coefficients, ϵ_i are uncorrelated random variables with zero mean and common variance. These regression parameters are specified by minimising the weighted least squares criterion given as follows:

$$\sum_{i=1}^{M}\{\theta_i - (\alpha + (\boldsymbol{s}_i - \boldsymbol{s})^t\beta)\}^2 K_\delta(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \quad \text{(C.3)}$$

The corrected particles θ_i^* can then be estimated using the regression model as follows:

$$\theta_i^* = \theta_i - (\boldsymbol{s}_i - \boldsymbol{s})^T\beta \quad \text{(C.4)}$$

The solution to the weighted least squares criterion to estimate α and β holds the following form:

$$(\alpha, \beta) = (X^T W X)^{-1} X^T W \theta \tag{C.5}$$

where
$$X = \begin{pmatrix} 1 & s_{11} - s_1 & \dots & s_{1q} - s_q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_{m1} - s_1 & \dots & s_{mq} - s_q \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \end{pmatrix} \tag{C.6}$$

where q is the dimension of the summary statistic and m is the number of parameters. W is an $m \times m$ diagonal matrix of the weights whose diagonal entries are $K_\delta(\|s_i - s\|)$. A summary of the steps followed in this algorithm is shown in algorithm C.1.

Algorithm C.1

---

1. Repeat the following until M points have been generated:
    a) Draw $\theta_i \sim \pi(\theta)$.
    b) Simulate $x_i \sim p(x|\theta_i)$.
2. Compute k_j, the empirical standard deviation of the $S_j(x)$.
3. Define $\rho\big(S(x), S(y)\big)$: $\sqrt{\sum_{j=1}^{s} \big(S_j(x)/k_j - S_j(y)/k_j\big)^2}$.
4. Choose tolerance δ such that the proportion of accepted points $P_\epsilon = N/M$.
5. Weight the simulated points $S(x_i)$, using $k_\epsilon\left(\rho\big(S(x_i), S(y)\big)\right)$ where
$$k_\epsilon(t) = \begin{cases} \delta^{-1}\,(1 - (t/\delta)^2) & t \le \delta \\ 0 \qquad\qquad x, & t > \delta \end{cases}$$
6. Apply weighted linear regression to the *N* points, to obtain an estimate of $E(\theta|S(x_i))$.
7. Adjust $\theta_i^* = \theta_i - E(\theta|S(x_i)) + E(\theta|S(y))$.
8. The θ_i^*, with weights $k_\delta\left(\rho\big(S(x_i), S(y)\big)\right)$, are drawn from the adjusted distribution.

Lintusaari et al. (2016) illustrated the local linear regression technique in a schema (Figure C.1). They stated that the same accuracy can be achieved with a higher tolerance value as that achieved with a lower one, if a regression analysis is applied to the sampled parameters, which increases the computation efficiency. The regression adjustments in this algorithm are done for one parameter at a time assuming that the parameters and their residuals are uncorrelated. Even if a multivariate regression was applied to all the parameters, the same results would be obtained since the residuals between parameters are assumed to be uncorrelated when applying the least-squares method to minimise the distances while fitting the model.

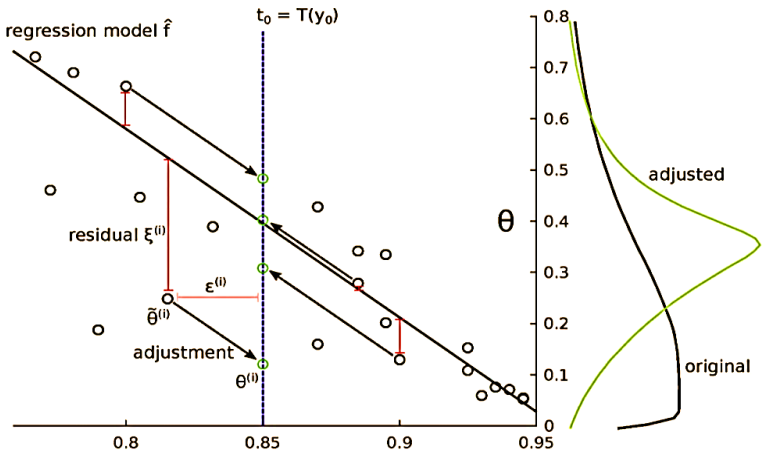

Figure C.1: Illustration of the local linear regression adjustments. t_0: observation summary statistic (Lintusaari et al. 2016)

In order to benefit from this technique, the fitted regression has to be accurate to avoid adjusting the particles in the direction of un-informative statistics. Moreover, it can easily happen that the adjusted particle falls outside the range of values assigned to the distribution, particularly if the slope of the regression is high. Thus, this technique should be applied carefully so that we do not end up with a distribution that is less accurate than the posterior PDF resulting from the ABC algorithm. Hickerson et al. (2006) transformed each particle that ended up having a value outside the range of its prior distribution after regression adjustment to the prior distribution boundary value.

This regression adjustment of the posterior distribution after the termination of the ABC algorithm is called post-processing. Different ABC-post-processing techniques and approaches have been proposed by different authors. Some of these approaches are regression-based methodologies and some follow a different concept.

Appendix D. Perturbation kernels

There exists multiple different kernel functions that can be adapted to the ABC sequential samplers presented in the thesis. Filippi et al. (2013) proposed an optimal covariance matrix:

$$\sum_{\theta^{(t-1)}}^{(t)} = E_{\theta^{(t)} \sim p(.|\mathcal{X})}\left[\left(\theta^{(t)} - \theta^{(t-1)}\right)\left(\theta^{(t)} - \theta^{(t-1)}\right)^{T}\right] \quad \text{(D.1)}$$

They recommended the use of such covariance matrix as a general rule of thumb as it yielded the highest acceptance rate between different other kernels in their toy examples under an acceptable computational cost.

An alternative to compute the covariance of the multivariate normal kernels using all the particles of the previous iteration, is to only consider the *K*-nearest neighbours of the particle in question. The reason behind the need of doing so is that considering all the particles might be inefficient if the parameters are highly correlated (Filippi et al. 2013). In other words, the correlation between the parameters is not all the time linear, there might exist different patterns in the correlation between two different parameters. In this case a covariance based on all the particles would not capture the local pattern and information needed around the particle to be perturbed. Thus, it is better to compute the covariance based only on the neighbour particles. The kernel is then a multivariate normal distribution having the particle in question as its centre and a covariance $\Sigma_{\theta,k}^{t}$ computed only from the *K*-nearest neighbours. The value of *K* in such kernels should be carefully chosen so that it is not too big in a way that it becomes similar to the standard multivariate kernel and not too small, so it leads to a narrow posterior distribution (no space exploration).

Appendix E. Complementary results

The estimation of the parameters not illustrated in the thesis are presented here. As a reminder, in chapter 5, five case studies are considered. The parameters estimation was evaluated using weighted Euclidean distance. The estimation of some parameters was visualised and convey clearly the behaviour of ARF. In this section all the parameters estimation for the five cases are presented.

E.1. Application of ARF on case 1

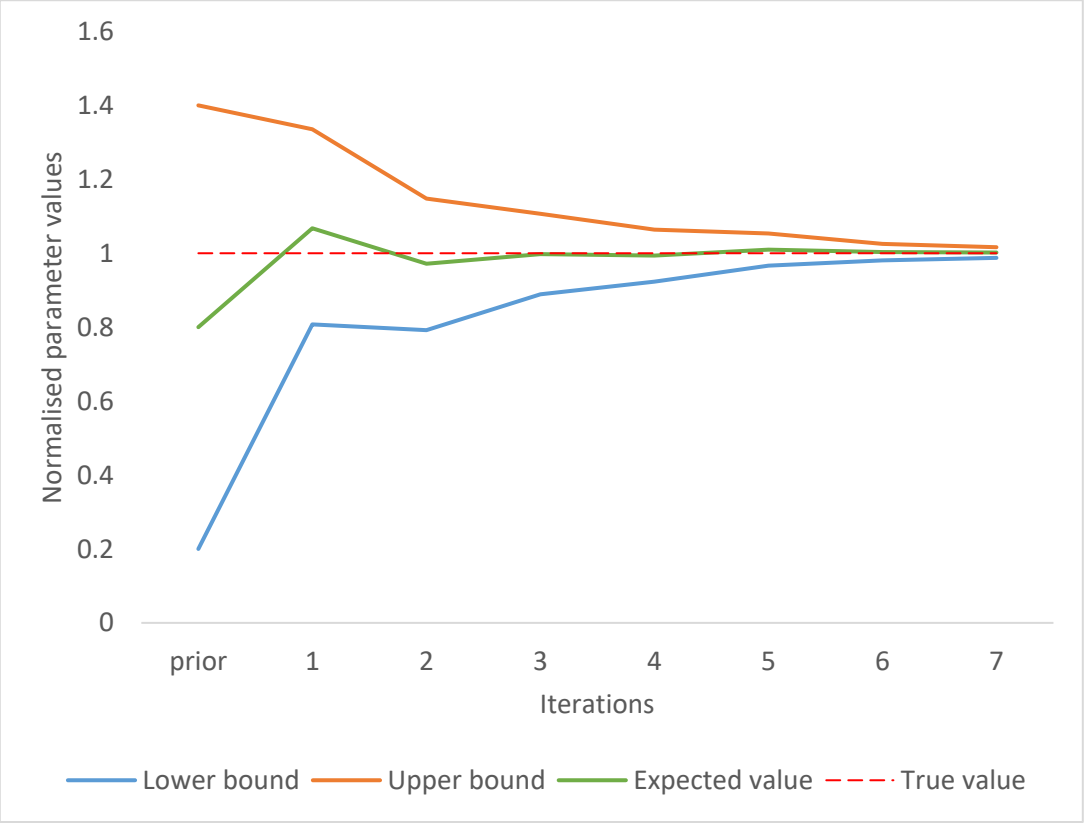

Figure E.1: Evolution of internal gains with ARF iterations (case 1)

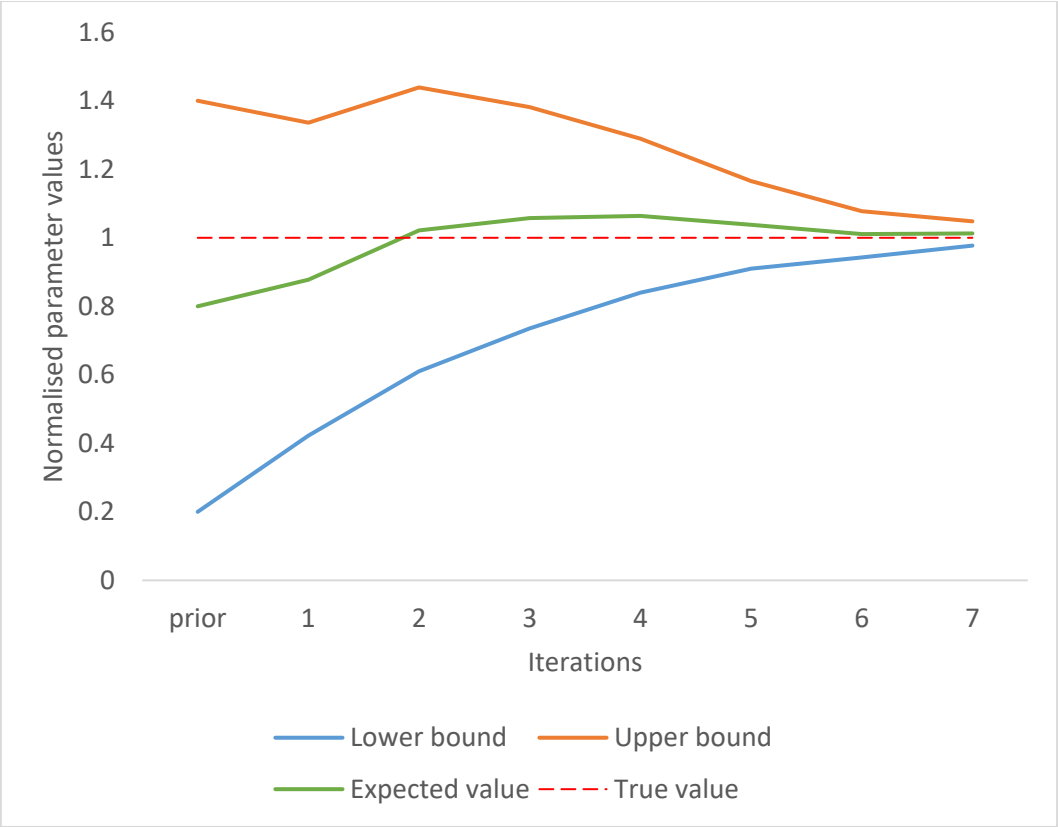

Figure E.2: Evolution of ventilation flowrate with ARF iterations (case 1)

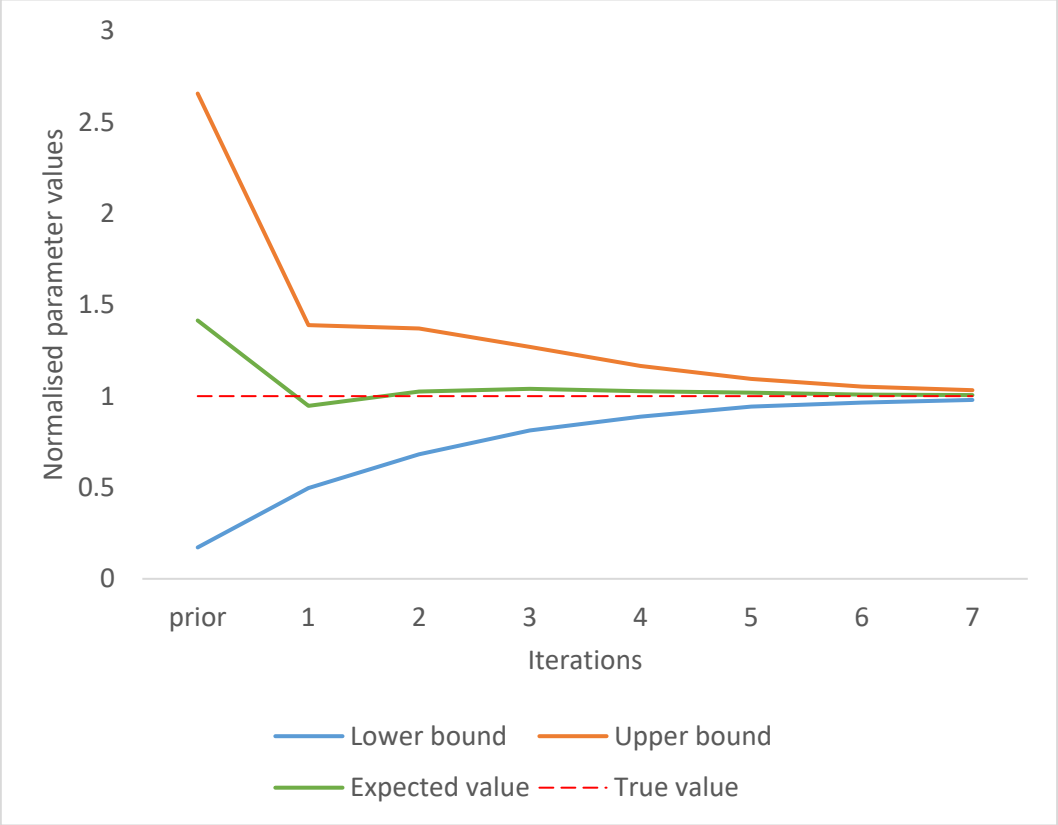

Figure E.3: Evolution of concrete specific heat with ARF iterations (case 1)

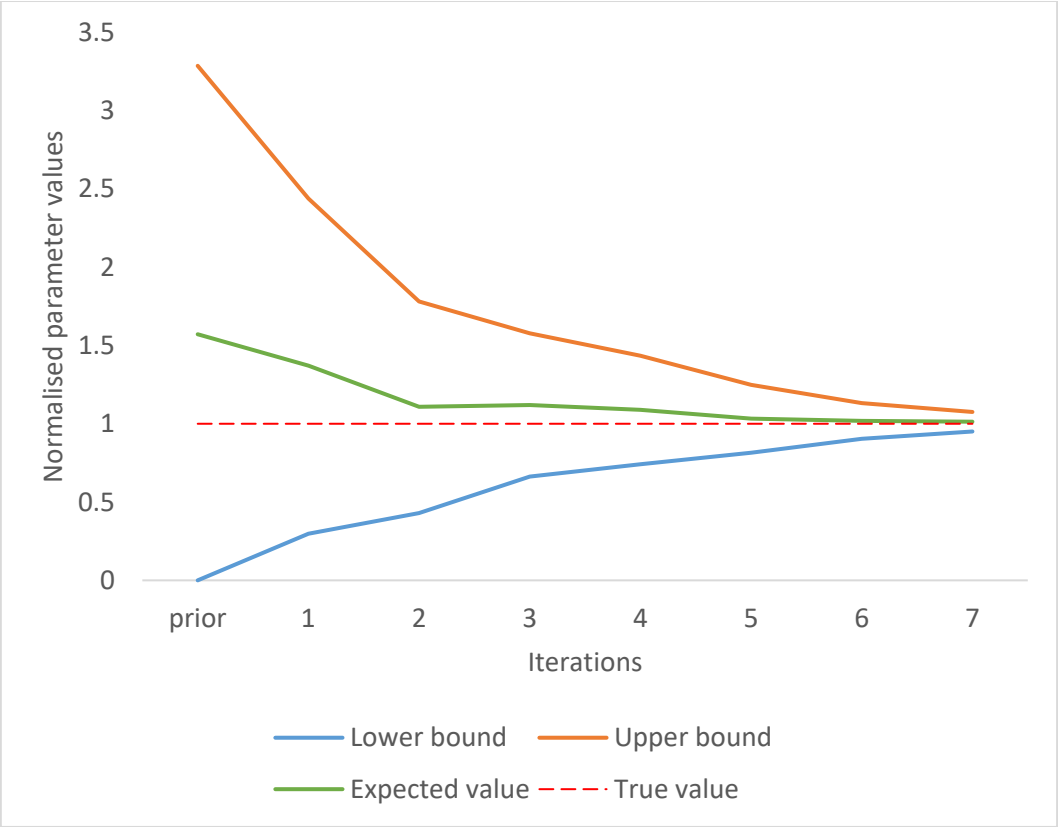

Figure E.4: Evolution of solar albedo with ARF iterations (case 1)

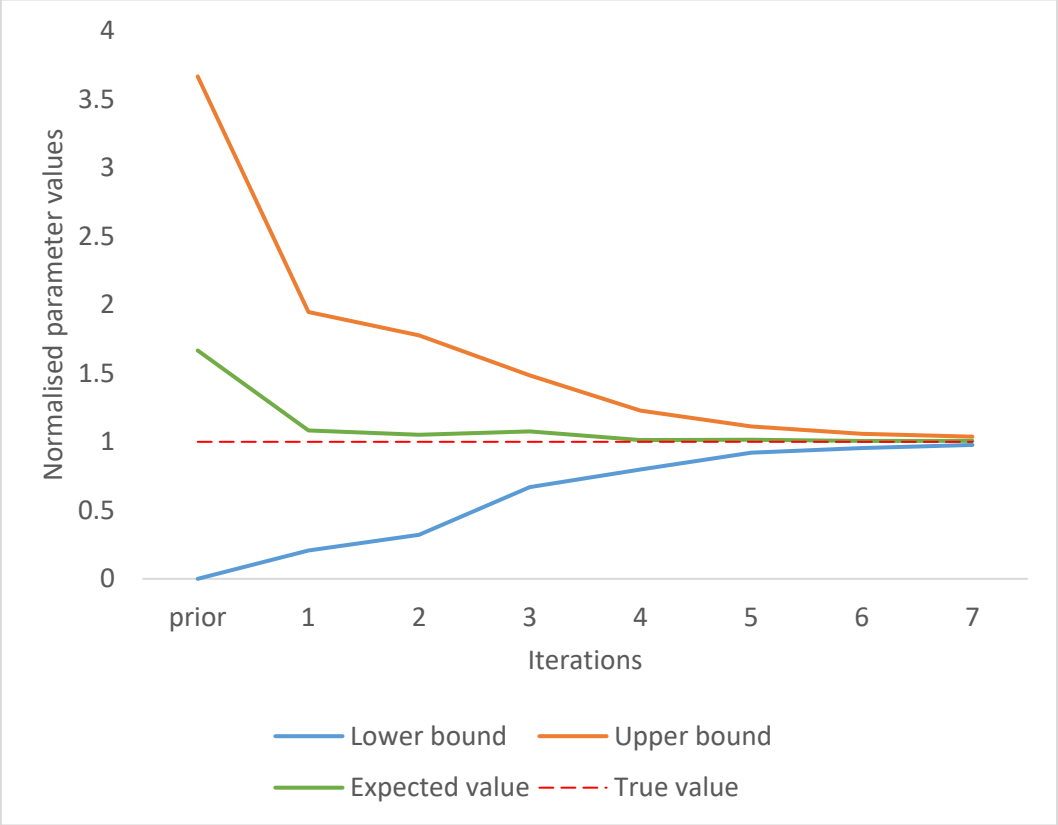

Figure E.5: Evolution of conducitivity of polystyrene with ARF iterations (case 1)

E.2. Application of ARF on case 2

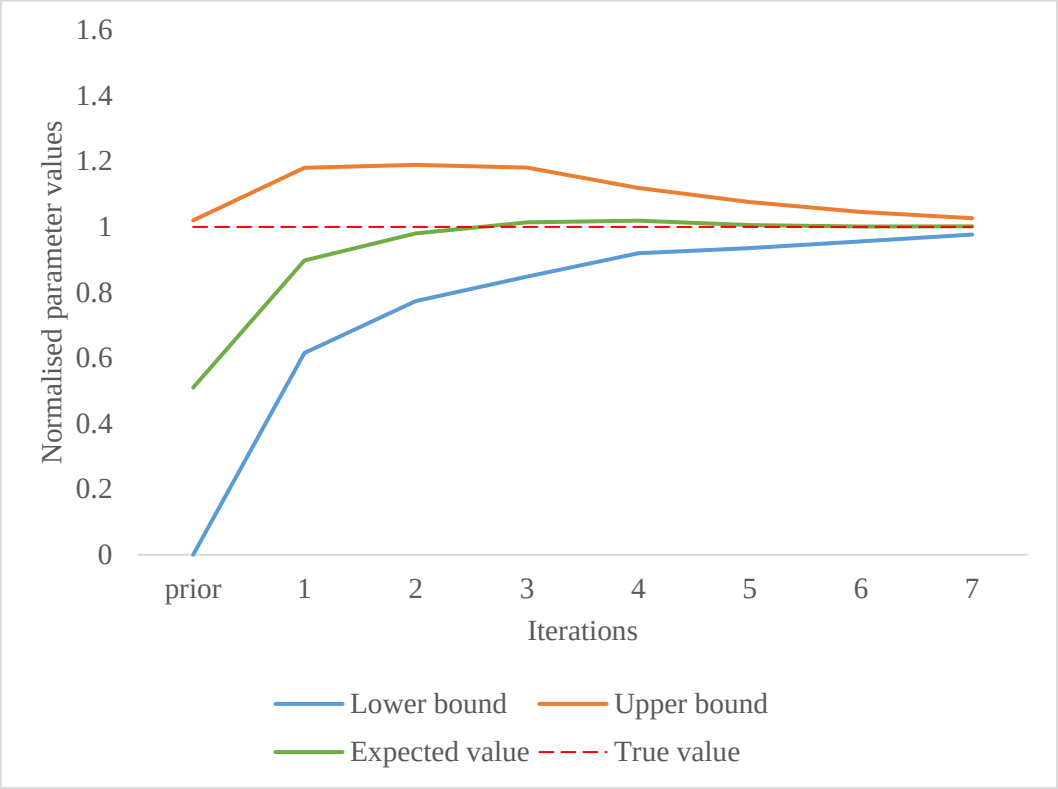

Figure E.6: Evolution of internal gains with ARF iterations (case 2)

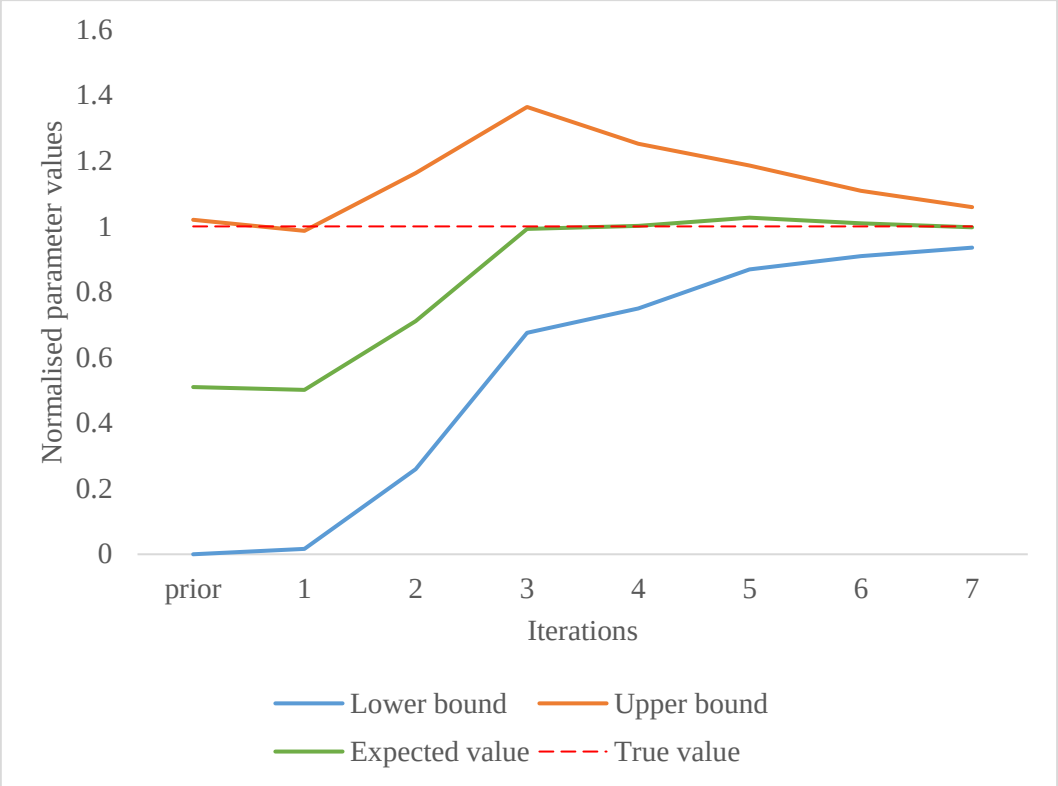

Figure E.7: Evolution of ventilation flowrate with ARF iterations (case 2)

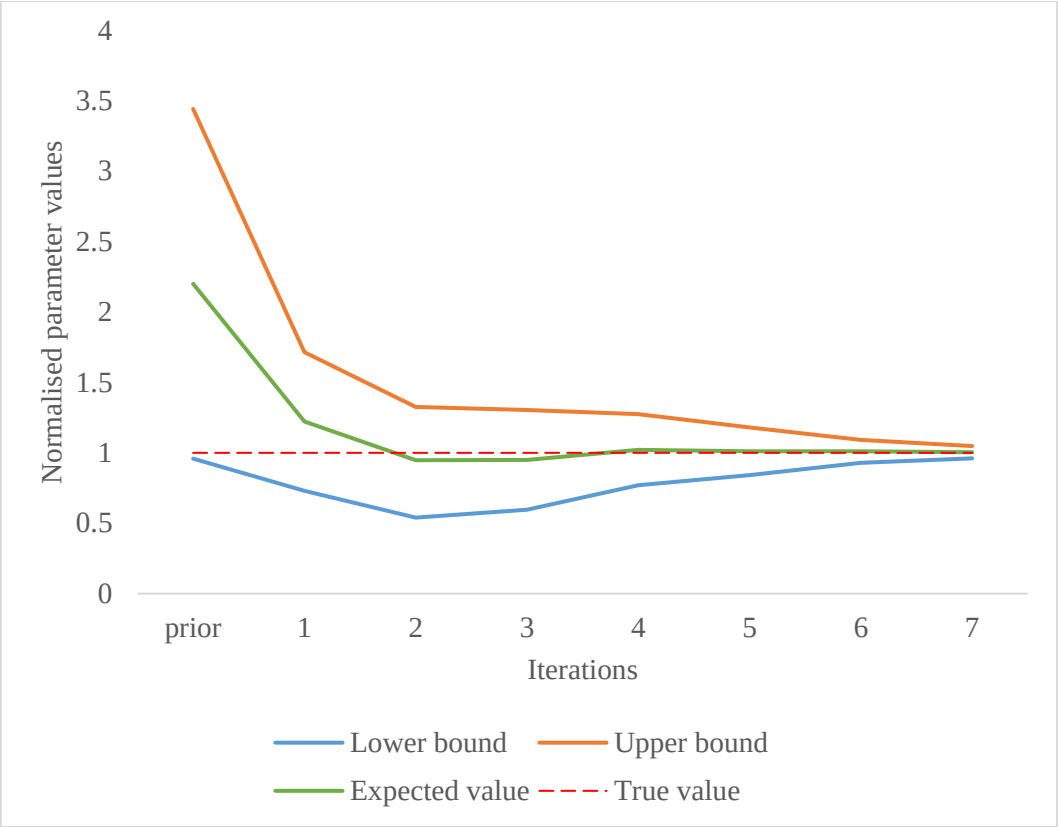

Figure E.8: Evolution of concrete specific heat with ARF iterations (case 2)

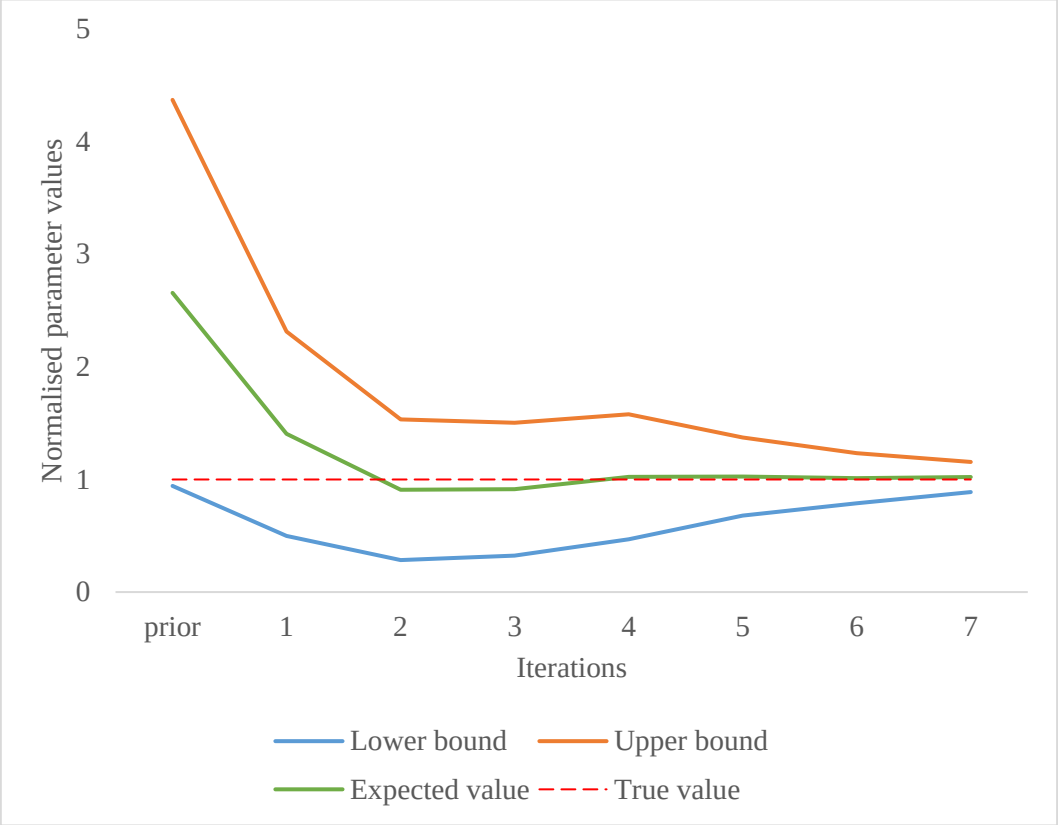

Figure E.9: Evolution of solar albedo with ARF iterations (case 2)

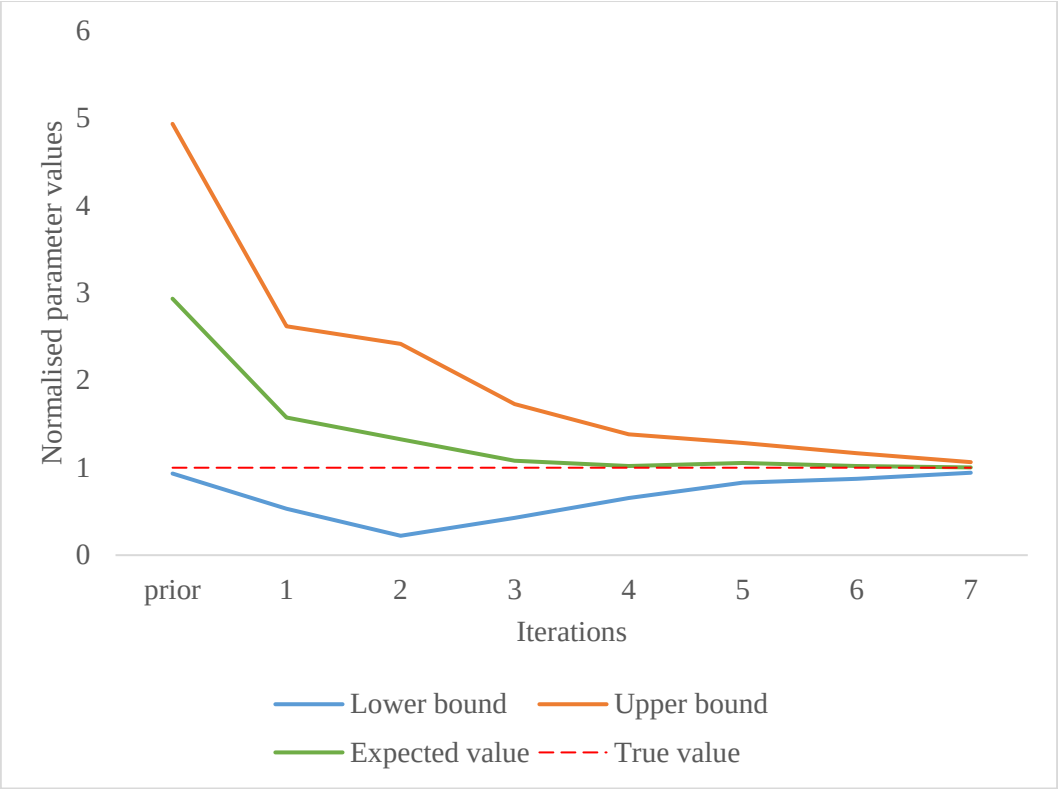

Figure E.10: Evolution of conductivity of polystyrene with ARF iterations (case 2)

E.3. Application of ARF on case 3

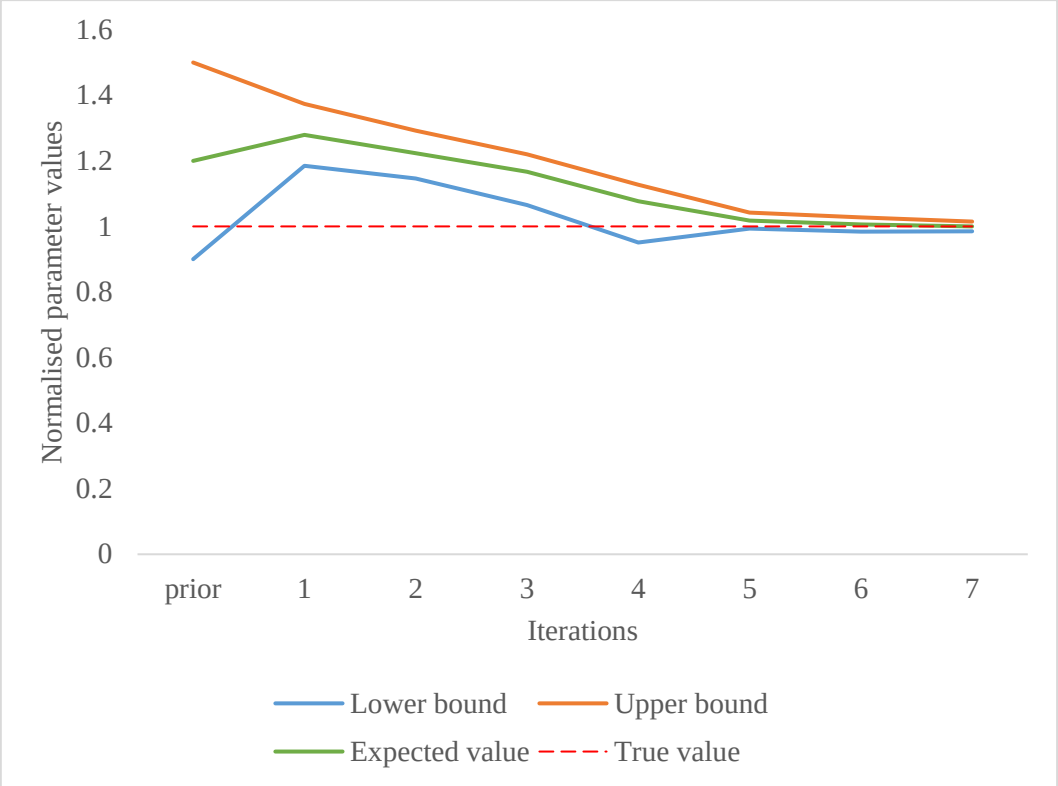

Figure E.11: Evolution of internal gains with ARF iterations (case 3)

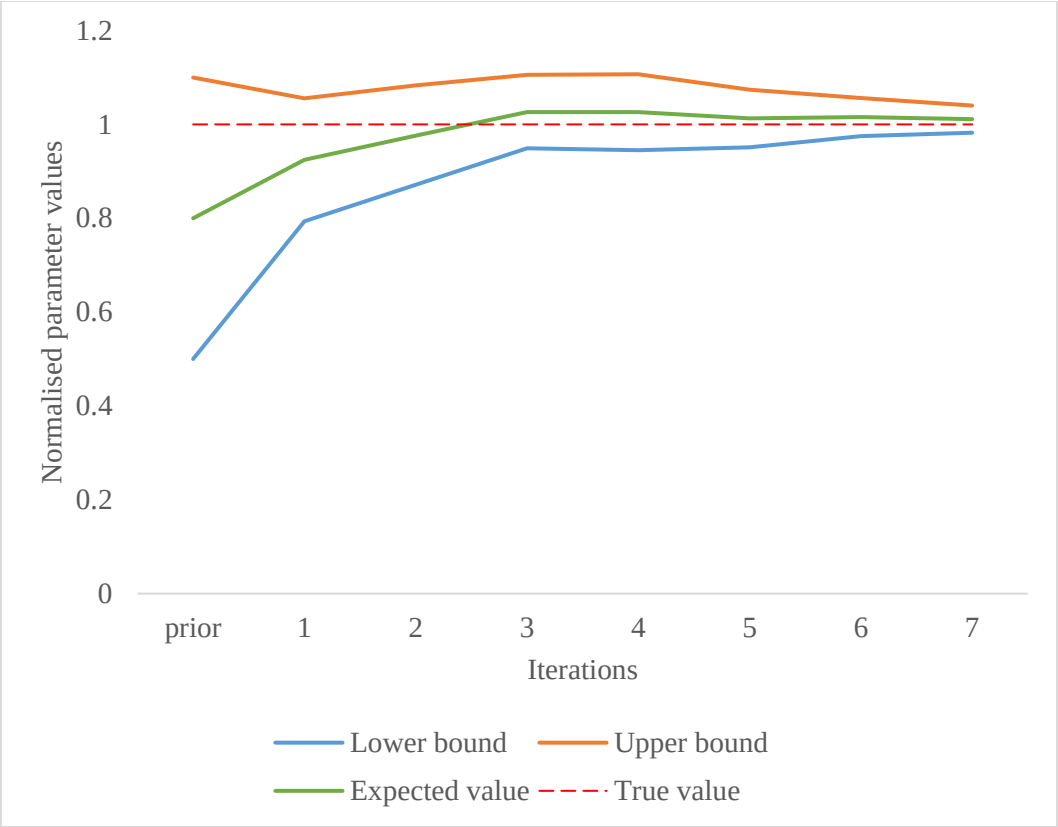

Figure E.12: Evolution of ventilation flowrate with ARF iterations (case 3)

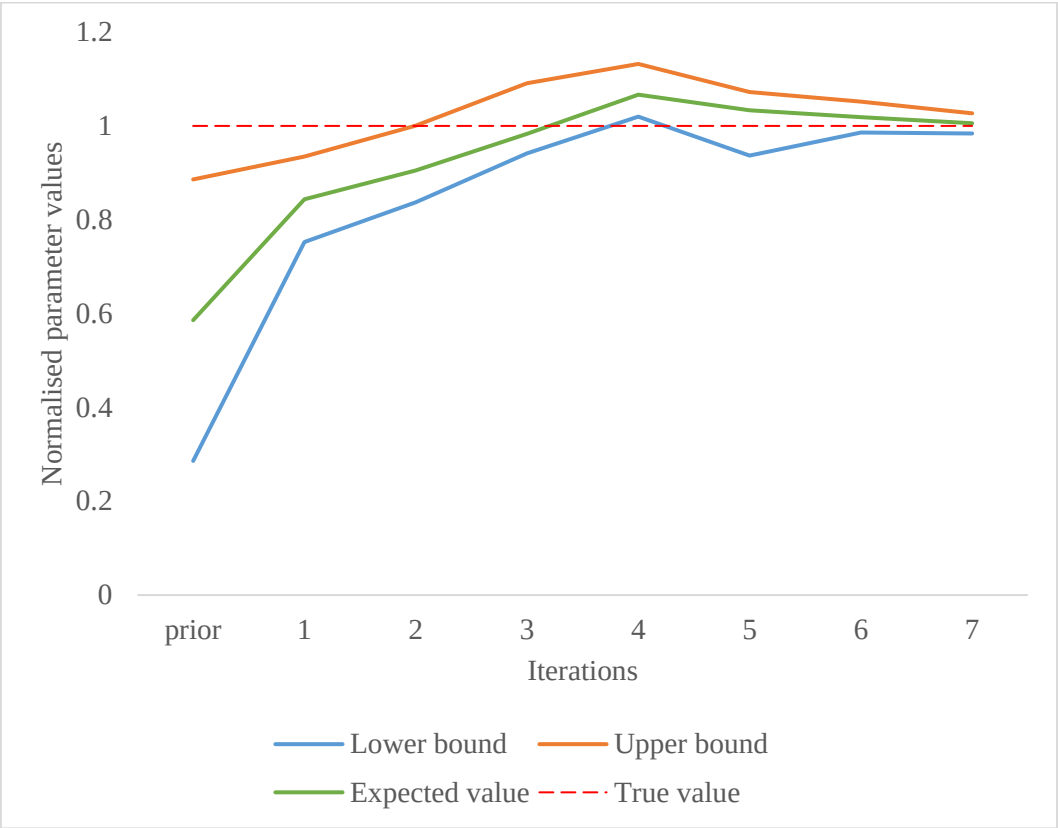

Figure E.13: Evolution of concrete specific heat with ARF iterations (case 3)

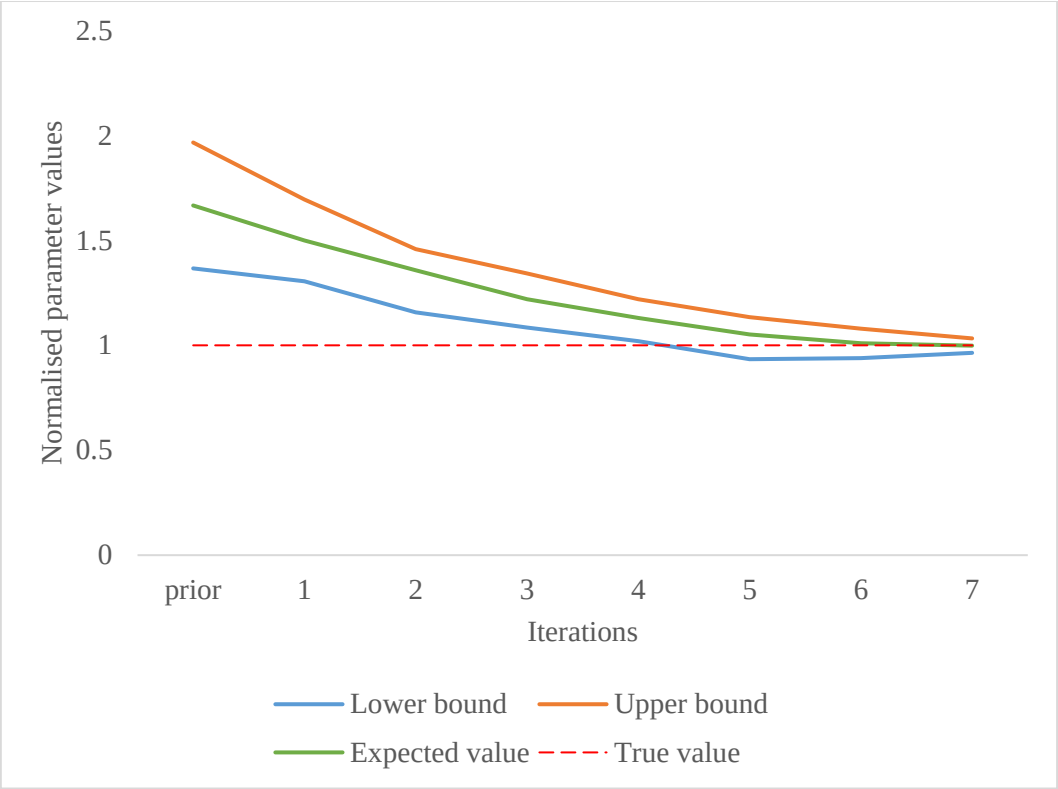

Figure E.14: Evolution of conductivity of polystyrene with ARF iterations (case 3)

E.4. Application of ARF on case 4

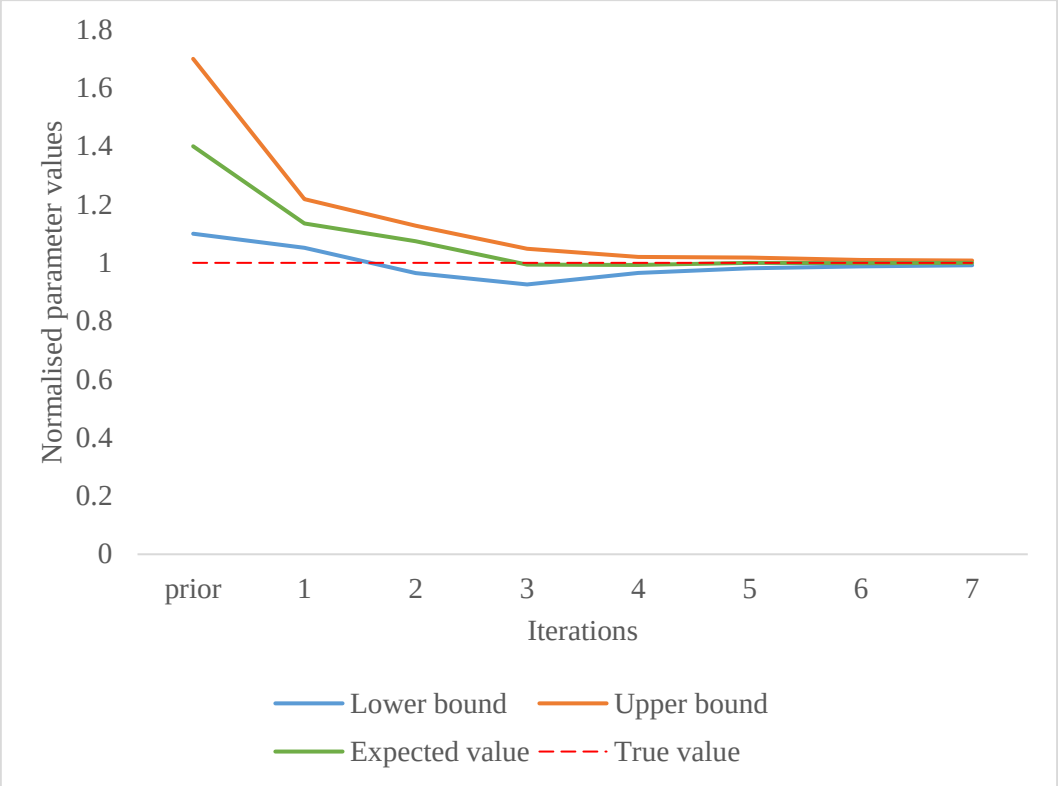

Figure E.15: Evolution of internal gains with ARF iterations (case 4)

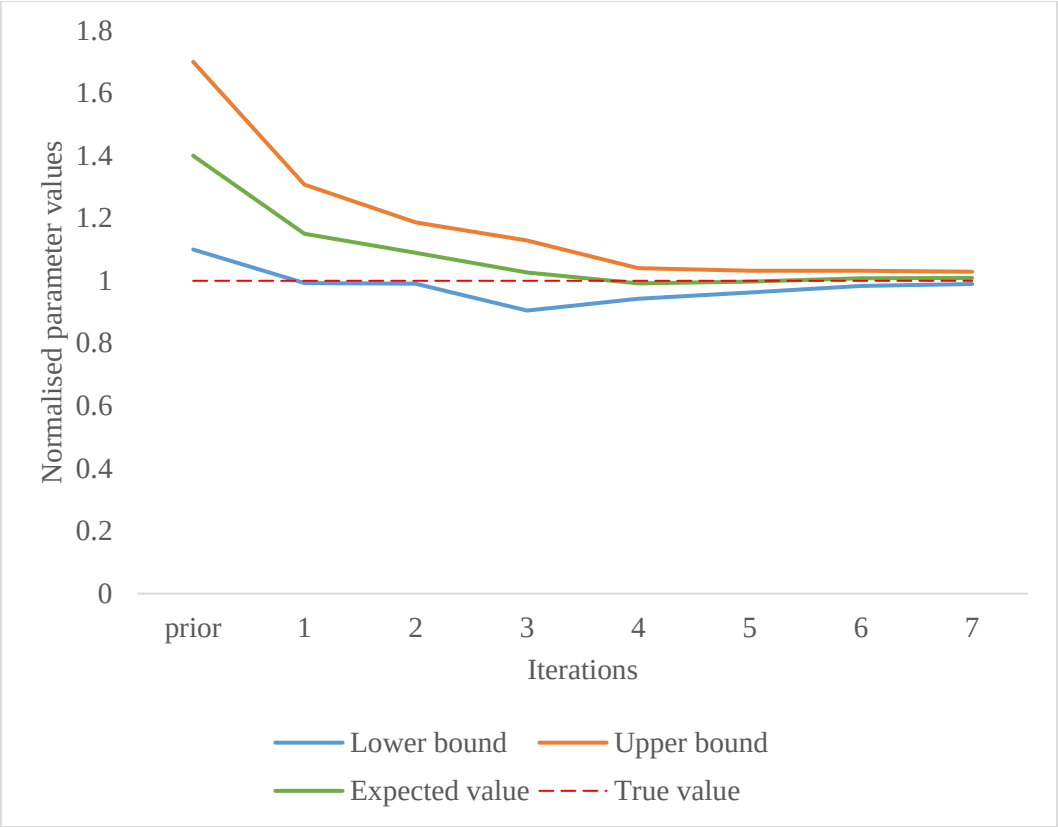

Figure E.16: Evolution of ventilation flowrate with ARF iterations (case 4)

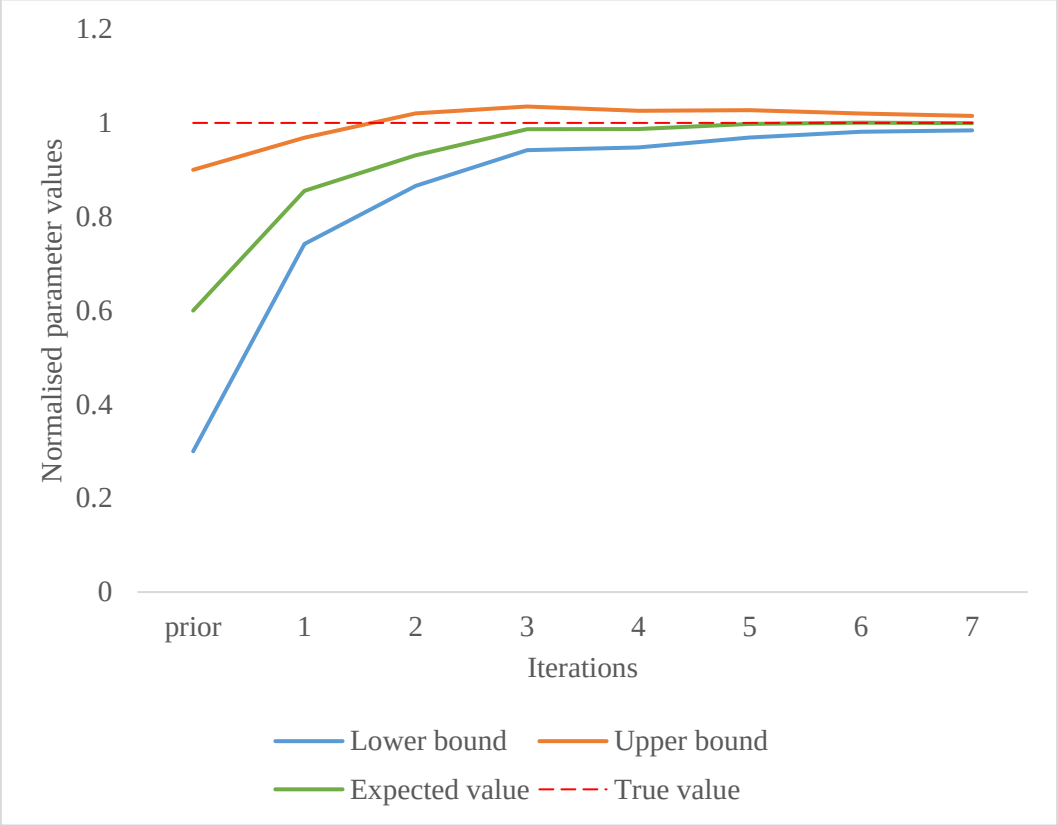

Figure E.17: Evolution of concrete specific heat with ARF iterations (case 4)

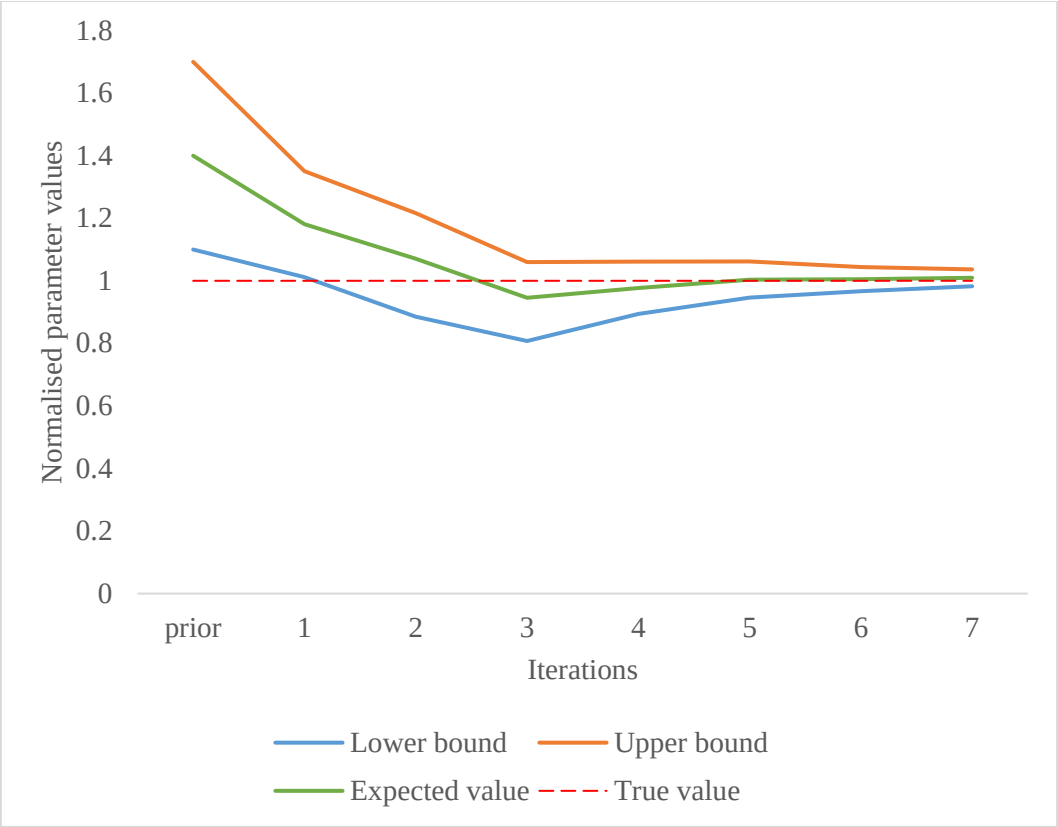

Figure E.18: Evolution of solar albedo with ARF iterations (case 4)

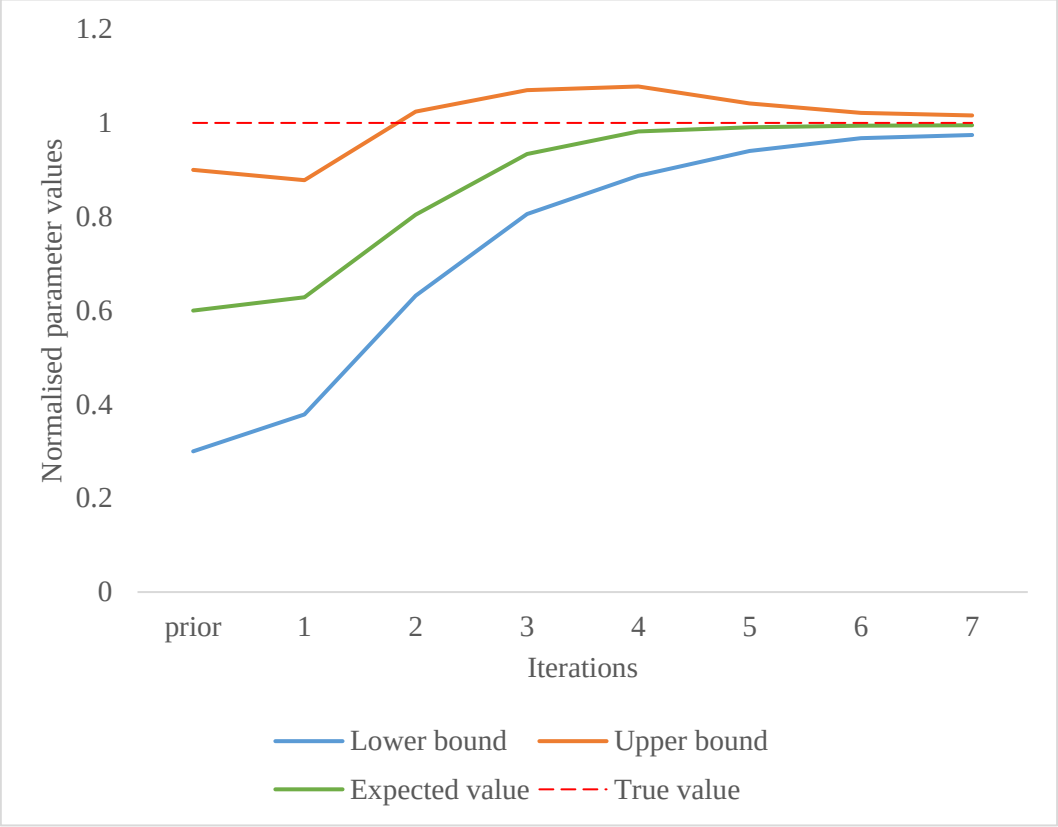

Figure E.19: Evolution of conductivity of polystyrene with ARF iterations (case 4)

E.5. Application of ARF on case 5

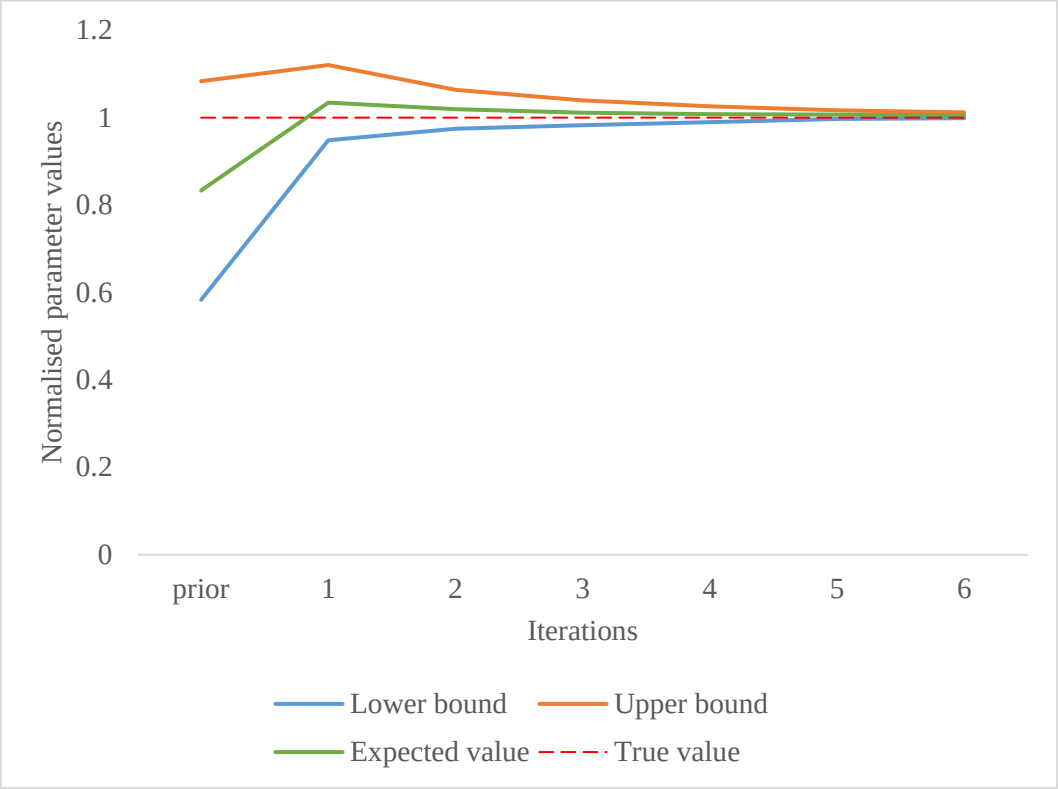

Figure E.20: Evolution of internal gains with ARF iterations (case 4)

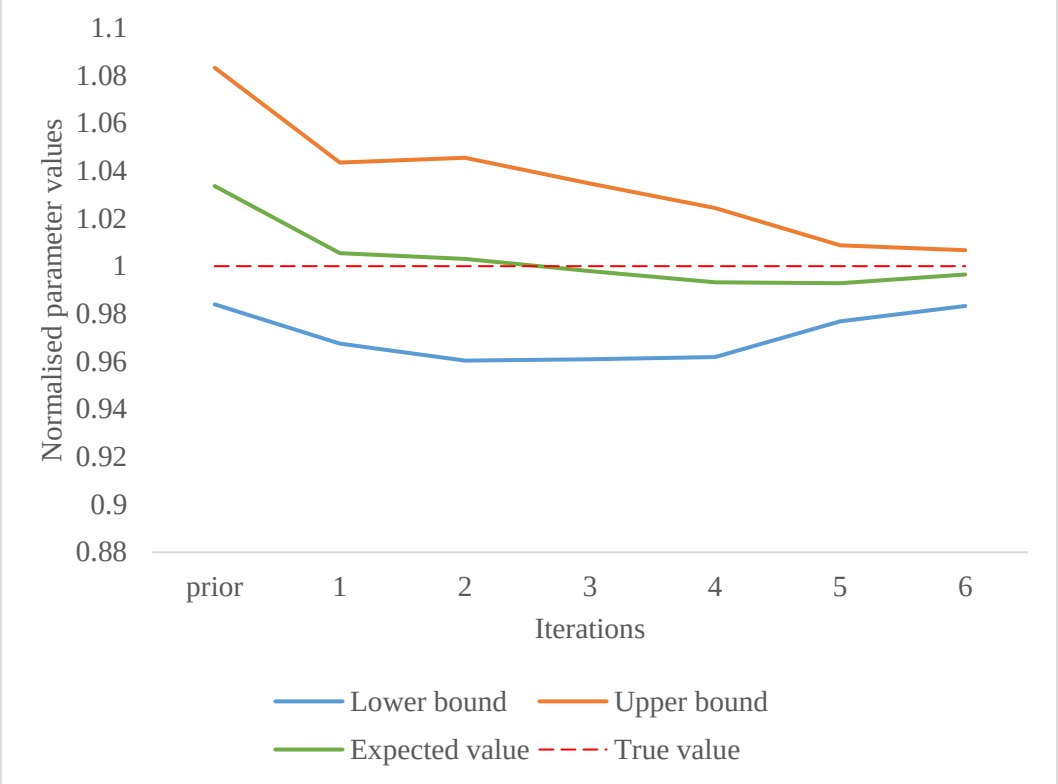

Figure E.21: Evolution of heating power with ARF iterations (case 4)

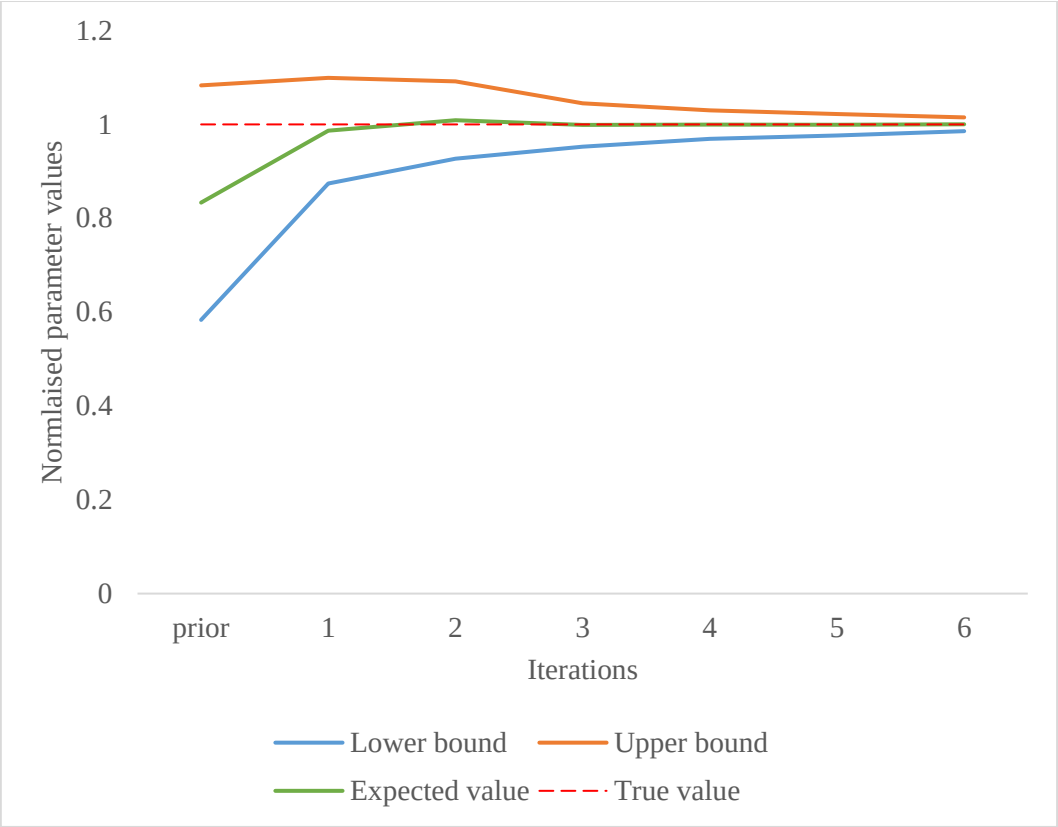

Figure E.22: Evolution of ventilation flowrate with ARF iterations (case 4)

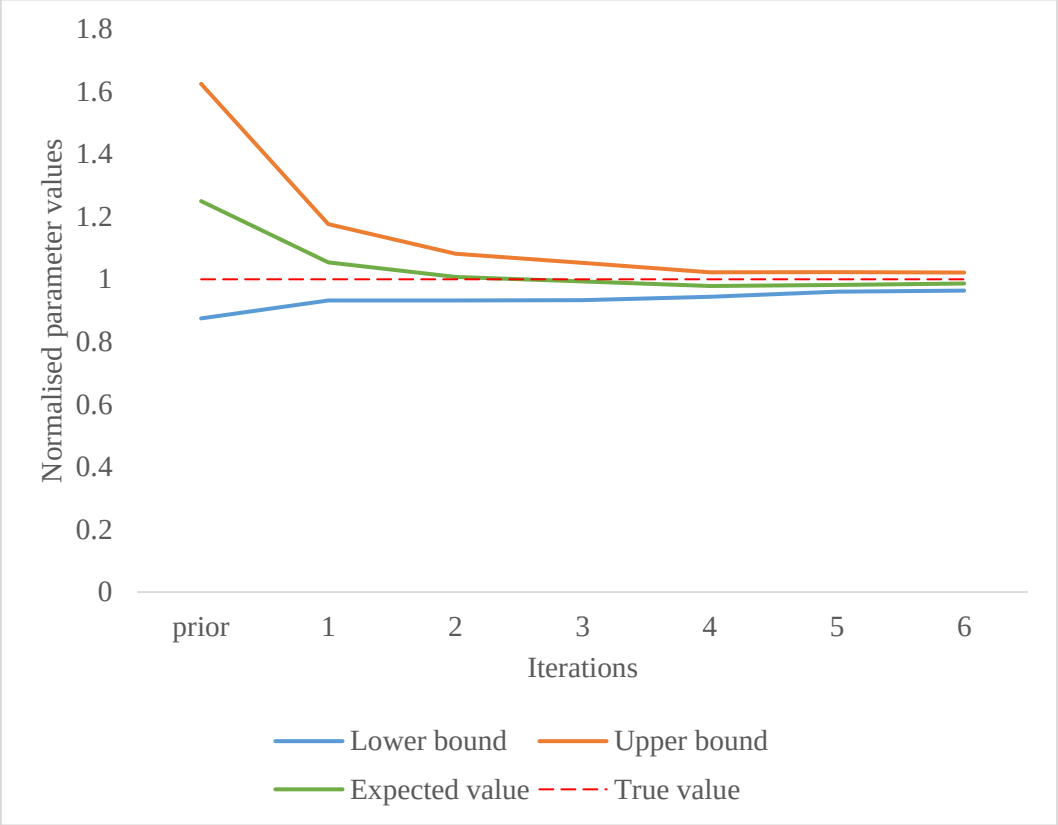

Figure E.23: Evolution of concrete specific heat with ARF iterations (case 4)

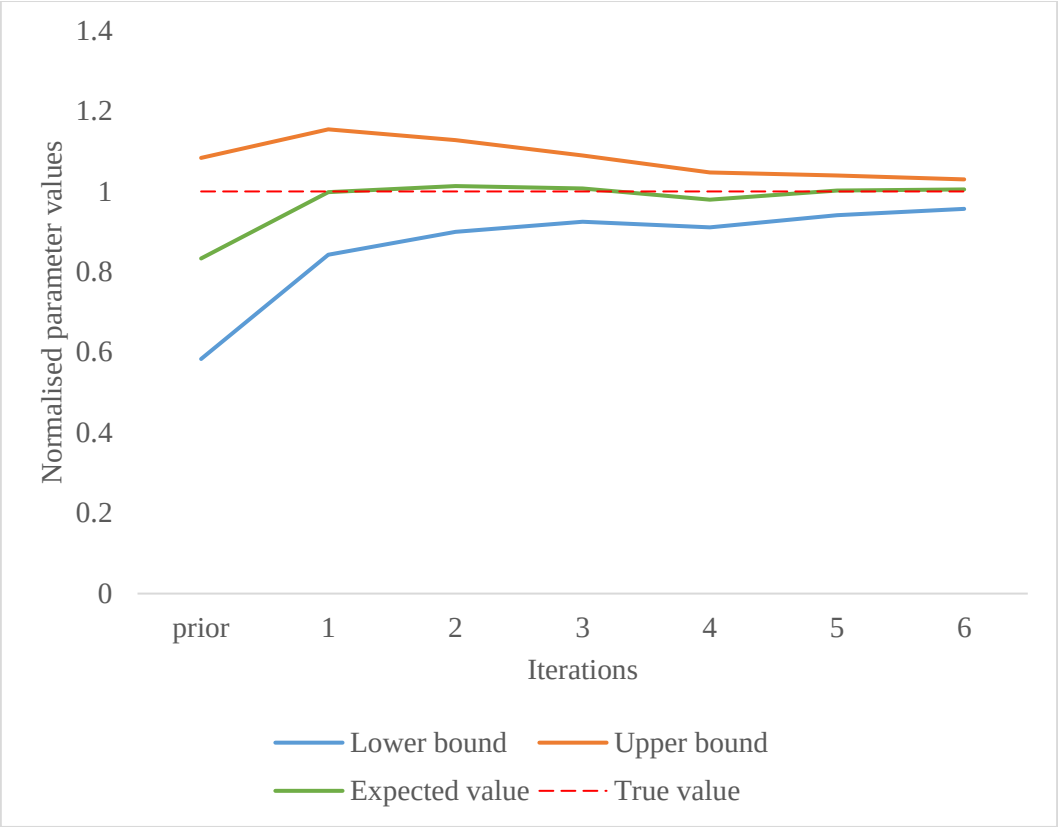

Figure E.24: Evolution of solar albedo with ARF iterations (case 4)

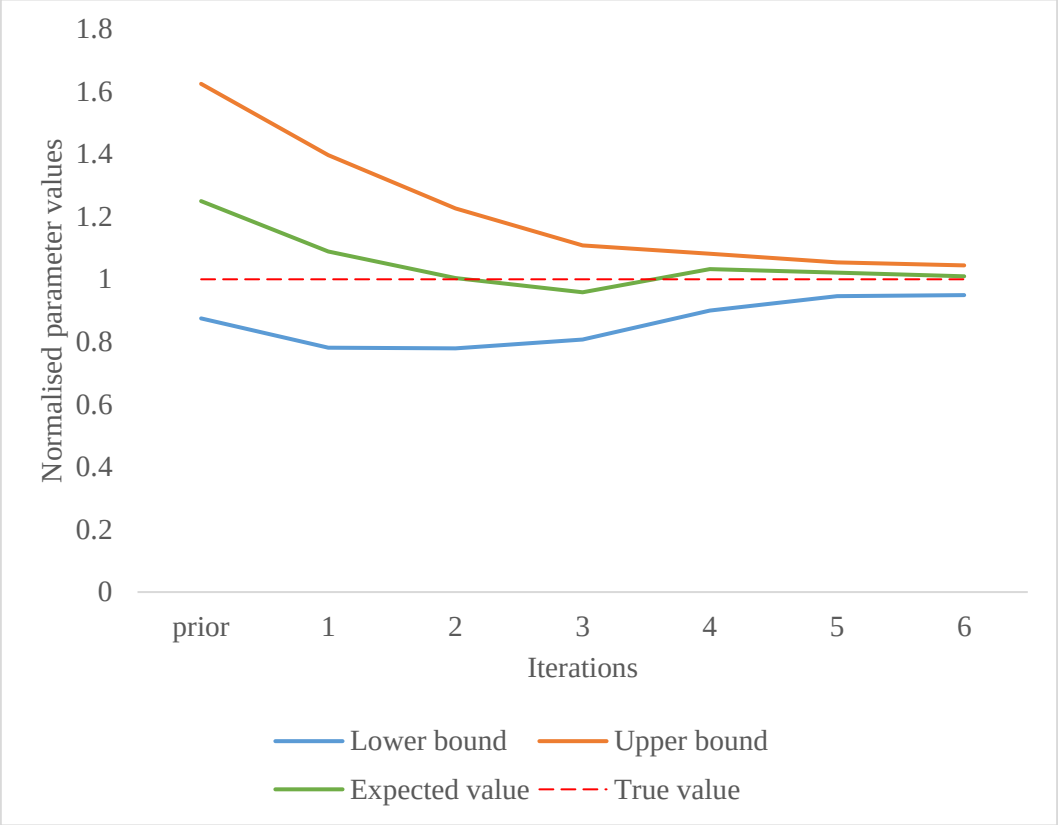

Figure E.25: Evolution of conductivity of polystyrene with ARF iterations (case 4)

ABSTRACT

Dynamic building energy simulation models are essential to analyse the energy performance of building renovation or new construction projects. However, these models are characterised by some degree of uncertainty and they could show poor fit to measured observations. Thus, calibration and uncertainty propagation have received an increasing attention in the field of building energy simulation. In this thesis, different Bayesian calibration methods are selected from literature and assessed in terms of accuracy and computational efficiency. A new method that is computationally faster than the ones found in literature is proposed and tested on virtual data. A detailed comparison between sensitivity analysis methods is conducted in terms of robustness, accuracy and computational efficiency using Sobol method as the reference method. Additionally, an identifiability analysis based on the sensitivity results is applied to rank the parameters not only in terms of importance but also to account for possible interactions. The effect of this step is evaluated in terms of calibration performance. Moreover, the choice of the number of parameters for calibration is studied on a virtual case study following an appropriate methodology. Finally, a real case study corresponding to real monitored data is used to check the findings of this thesis.

KEYWORDS

Sensitivity analysis, identifiability analysis, Bayesian calibration, random forest

RÉSUMÉ

Les outils de simulation énergétique dynamique des bâtiments sont essentiels pour analyser la performance de projets de rénovation ou de construction neuve. Cependant, ces modèles sont caractérisés par un degré d'incertitude et un biais est généralement constaté par rapport aux observations mesurées. Ainsi, le calibrage des modèles et la propagation des incertitudes ont reçu une attention croissante dans le domaine de la simulation énergétique des bâtiments. Dans cette thèse, différentes méthodes bayésiennes sont sélectionnées dans la littérature et évaluées en termes de précision et d'efficacité de calcul. Une nouvelle méthode plus rapide en termes de calcul que celles trouvées dans la littérature est également proposée et testée sur des données virtuelles. Une comparaison détaillée entre des méthodes d'analyse de sensibilité est effectuée en termes de robustesse, de précision et d'efficacité de calcul. De plus, une analyse d'identifiabilité basée sur les résultats de sensibilité est menée pour classer les paramètres non seulement en termes d'importance mais aussi pour tenir compte d'éventuelles interactions. L'effet de cette étape est évalué en termes de performance du calibrage. De plus, le choix du nombre de paramètres pour le calibrage est étudié sur une étude de cas virtuelle suivant une méthodologie appropriée. Enfin, une étude de cas réel correspondant à des données réelles monitorées est utilisée pour vérifier les conclusions de cette thèse.

MOTS CLÉS

Analyse de sensibilité, analyse d'identifiabilité, calibrage bayésienne, forêt aléatoire