Energy & Buildings 224 (2020) 110236

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Analysing uncertainty in parameter estimation and prediction for grey-box building thermal behaviour models ☆

O.M. Brastein [a,*], A. Ghaderi [b], C.F. Pfeiffer [a], N.-O. Skeie [a]

[a] Department of Electrical Engineering, Information Technology and Cybernetics, University of South-Eastern Norway, N-3918 Porsgrunn, Norway
[b] Department of Mathematics and Science Education, University of South-Eastern Norway, N-3918 Porsgrunn, Norway

ARTICLE INFO

Article history:
Received 21 December 2019
Revised 16 April 2020
Accepted 10 June 2020
Available online 23 June 2020

Keywords:
Thermal network models
Grey-box models
Profile likelihood
Bayesian parameter estimation
Markov Chain Monte Carlo
Parameter distribution
Posterior predictive distribution
Parameter identifiability

ABSTRACT

The potential reduction in energy consumption for space heating in buildings realised by the use of predictive control systems directly depends on the prediction accuracy of the building thermal behaviour model. Hence, model *calibration* methods that allow improved prediction accuracy for *specific* buildings have received significant scientific interest. An extension of this work is the potential use of calibrated models to estimate the *thermal properties* of an existing building, using measurements collected from the actual building, rather than relying on building specifications.

Simplified thermal network models, often expressed as *grey-box* Resistor-Capacitor circuit analogue models, have been successfully applied in the prediction setting. However, the use of such models as *soft sensors* for the thermal properties of a building requires an assumption of *physical interpretation* of the estimated parameters. The parameters of these models are estimated under the effects of both *epistemic* and *aleatoric* uncertainty, in the model structure and the calibration data. This uncertainty is propagated to the estimated parameters. Depending on the model structure and the dynamic information content in the data, the parameters may not be *identifiable*, thus resulting in *ambiguous* point estimates.

In this paper, the Profile Likelihood method, typical of a *frequentist* interpretation of parameter estimation, is used to diagnose parameter *identifiability* by projecting the likelihood function onto each parameter. If a Bayesian framework is used, treating the parameters as random variables with a probability distribution in the parameter space, *projections* of the posterior distribution can be studied by using the Profile Posterior method. The latter results in projections that are *similar* to the *marginal distributions* obtained by the popular Markov Chain Monte Carlo method. The different approaches are applied and compared for five experimental cases based on *observed* data. Ambiguity of the estimated parameters is resolved by the application of a *prior distribution* derived from a priori knowledge, or by appropriate modification of the model structure. The posterior predictive distribution of the *model output predictions* is shown to be mostly *unaffected* by the parameter non-identifiability.

1. Introduction

1.1. Background

The reduction of anthropogenic CO_2 emissions is perhaps the most important task in modern science. The energy consumed by space heating in buildings is considerable [1]. According to the Energy Performance of Buildings Directive (EPBD) [2], the energy consumed by buildings accounts for 40% of the total energy consumption within the European Union (EU). Hence, the development of model predictive control strategies that can effectuate energy reductions by improved thermal control has received significant scientific interest [3,4]. For control systems, development of accurate *prediction* models is essential.

Another application of interest for building thermal modelling is the *classification* of building properties related to space heating, for improved evaluation of the energy *performance* of existing buildings [5]. By classifying actual energy performance, development of taxation schemes could be utilised to motivate investments in energy reduction technology. Given that there is often discrepancies between physical buildings and their blueprints, typically due to continuous modifications or workmanship issues, energy classification schemes could with benefit be based on energy and temperature data recorded from the building to be

☆ This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

* Corresponding author.
E-mail address: ole.m.brastein@usn.no (O.M. Brastein).

https://doi.org/10.1016/j.enbuild.2020.110236
0378-7788/

classified. A popular method for modelling building thermal behaviour is the use of simplified thermal network models expressed as a Resistor-Capacitor analogue [6,7,4,3,8,9]. Regardless of their proven efficiency in the prediction setting, the parameters of a thermal network model may not be suitable as *soft sensors* for monitoring building properties, since this assumes a physical interpretation of the parameters as constants of the physical building [5,10]. For such an assumption to be justified, the verification of *parameter identifiability* is essential, in order to ensure unambiguous parameter estimation.

1.2. Previous work

1.2.1. Thermal behaviour models of buildings

For widespread use of model predictive control and/or classification systems in buildings, a simple modelling method that can both produce *physically interpretable* parameters *and* make accurate *predictions* of future thermal behaviour is needed. Models of building thermal behaviour based on exact physical specifications of a building often become intractable due to the complexity of building structures, and may require specialised software to simulate [1]. Additionally, existing buildings may deviate from blueprints and specifications of the building, which further exacerbates the challenge of developing a physics-based *white-box* model [5]. In contrast, data-driven models typically use simple model structures, with parameters that are calibrated from data acquired from existing buildings. Such data-driven *black-box* models, e.g., from system identification methods, typically have improved prediction performance due to being calibrated for specific buildings, but in general lack physical interpretability [11–15].

A reasonable compromise between the physics-based white-box and the data-driven black-box models is the use of *grey-box* thermal network models [3,4,7,9,5]. Thermal network (TN) models, typically expressed as Resistor-Capacitor (RC) electric analogue models, are based on a *naive* physical, *cognitive* understanding of the building thermodynamics, with relatively few *lumped* parameters that are calibrated from observational data. These models contain significant *epistemic* uncertainty in their formulation, resulting from model approximations and unmodelled or unrecognised disturbances [8], in addition to the *aleatoric* uncertainty introduced by random measurement noise. Hence, they can with advantage be formulated using stochastic differential equations (SDE) [16,17,8,18,19]. Since the structure of a grey-box TN model is developed based on a physical description of the building, the parameters are often assumed to have a physical interpretation. However, due to the inherent uncertainties involved in the formulation of such models, a through analysis of parameter identifiability, which may lead to ambiguous parameter estimates, is needed prior to such interpretation.

Another point of interest regarding interpretation of TN models is the model *states*. The temperature *state nodes* in the RC circuit are typically chosen to representation a specific part of the building, e.g., the room interior or the building envelope internal surface, hence a physical interpretation of the states are assumed from the model structure. However, since the parameters that determine the *relationship* between these nodes are calibrated from measured data, the model is trained to predict the temperature at the specific sensor locations [20]. If the states are directly measurable, each state corresponds to a specific sensor location. Hence, the physical interpretation of TN model states is determined by both the model structure and the sensor location. Compared to the black-box system identification (SSID) paradigm, where the model structure to be calibrated is some *general* state mapping, the grey-box TN structure *constrains* the state representation. in comparison for an SSID model, which also effectually learns to predict the system response at the sensor locations, a change of *basis* for the state space will result in equivalent descriptions of the system, with the same outputs given the same measured data, but with *different* state representations.

1.2.2. Parameter estimation

Estimation of parameters requires a well-defined objective function. Using a *statistically founded* objective function, such as the *likelihood* function or the *posterior distribution,* computed from Bayes' theorem by the inclusion of a *prior* distribution, of the parameters, allows the use of statistical tools for model validation and analysis [16,10,21]. The *evaluation* of the likelihood function and/or the posterior parameter distribution for SDE models has previously been presented in detail in the Continuous Time Stochastic Modelling (CTSM) framework [16,22]. By utilising a Kalman Filter (KF) to compute the one-step ahead prediction *residuals,* which are subsequently assumed normally distributed, the likelihood can be efficiently evaluated for an SDE model [16]. The grey-box SDE approach has been claimed as a natural framework for modelling dynamic systems in general [23].

1.2.3. Parameter identifiability analysis and prediction accuracy

A common assumption for parametrised model structures is that there exist an *unambiguous* set of parameters, which is optimal in the sense that it produces the best *model fit* in some specified statistical sense. However, there are cases for which the objective function used for the estimation of parameters is in some way *non-informative* for a subset of the parameters, thus resulting in *ambiguous* solutions. This subset of parameters is denominated as *non-identifiable*. If the *non-identifiable* parameters are perturbed in some way, the objective function is either *unchanged*, or the change is *insufficient* to determine the bounds of the estimated parameter with a desired prescribed level of *confidence* [10]. A good diagnostic tool is found in the framework of the *Profile Likelihood* (PL) method [10,21,24,5].

Since the objective function compares model predictions with measured data, non-identifiability may be caused by either the *model structure* or by a lack of *dynamic information* in the data. The former is the cause of *structural* non-identifiability, which presents as a flat equipotential *manifold,* bounded or unbounded depending on the model structure, in the parameter space [10]. Structural identifiability is well covered in the literature, and there exist several diagnostic methods based on a multitude of theoretical foundations [25,26,10,21,13].

If non-identifiability results from a lack of *dynamic information* in the calibration data, the affected parameters are diagnosed as *practically* non-identifiable. For a parameter to be identifiable according to the PL method [10], the *likelihood-based confidence interval* (CI), and subsequently also the likelihood profile, must be bounded in *both* directions. Hence a *practically* non-identifiable parameter may be diagnosed by inspecting the likelihood profile for the presence of a well-defined optimum that is *insufficiently pronounced* to produce a bounded CI [10].

The PL method, based on the likelihood function and computation of CIs, has a distinctly *frequentist* approach to parameter estimation. If a *Bayesian* framework is used, where parameters are treated as *random variables* that have a distribution in parameter space, the *Markov Chain Monte Carlo* (MCMC) method [21,27–31] can be used to *infer* the posterior distributions from the measurement data, typically visualised by obtaining marginal posterior distributions for single parameters or pairs of parameters [30,31,27]. The Bayesian framework combines the likelihood function with a *prior* by use of Bayes' theorem, thus computing the posterior distribution of the parameters [27]. The use of the Bayesian framework and MCMC for calibration of TN models was also reported in [32]. Alternatively, a variation of the PL method, called the *Profile*

Posterior (PP) method [21], may be used to visualise the posterior distribution by obtaining projections, rather than marginal distributions. Similar arguments w.r.t. the identifiability of parameters drawn from the PL method can be applied to the posterior distribution [21].

There are also several other methods that can be used to investigate parameter identifiability, some of which are reviewed in [33]. Some possibilities are the use of the Hessian matrix evaluated at the optimal estimate to compute confidence bounds, and the testing for convergence problems in the optimisation algorithm by repeated optimisations with randomized initial guess [34]. For simple linear models, structural identifiability can sometimes be evaluated analytically [35]. Another possibility is the use of graphing tools to analyse the interactions between parameters and model output [36].

Since the parameter non-identifiability results from the objective function being *non-informative* for a sub-set of the parameters, adding*more information* to the estimation problem is a reasonable strategy towards resolving the non-identifiability. Experimental redesign may be used in order to collect more *informative* data, either by *improved* dynamic information content in existing measurements or by adding *new* measurements from the system $\mathcal{S}$ [21,10]. The literature on system identification covers a range of experimental design considerations, including optimal experimental design for certain types of systems, see e.g. [11,37]. A popular approach is the use of a Pseudo Random Binary Sequence (PRBS) applied to the actuator which may result in improved system excitation, thus improving practical identifiability of model parameters [6]. However, for occupied buildings, the choice of excitation for the active heating system may be limited due to occupant demands. If obtaining more data is not possible, redesigning the model structure $\mathcal{M}$, such that the model better represents the actual experimental data collected, may also resolve the non-identifiability [21,10].

Finally, an important observation is that model structures with non-identifiable parameters can also provide reasonable predictions of the system outputs, but the non-identifiable parameters are arguably without a physical interpretation and can be considered *nuisance* parameters [5]. Indeed, ambiguous parameters without physical interpretation is the *norm* in traditional black-box calibration methods, such as system identification (SID) [14,15,11–13].

1.3. Overview of paper

In this work, the two projection-based methods, PL and PP, are compared to the MCMC method, on the basis of five experimental cases with differences in model structure, use of priors, identifiability of parameters and choice of training data. The theoretical foundation for the methods is presented in Section 2. The model, data and experimental setup of each case is presented in Section 3. The results are presented and discussed in Section 4, and the work concluded in Section 5.

2. Methods

2.1. Overview

In the sequel, the *Profile Likelihood* (PL) and the *Profile Posterior* (PP) methods [10,21] are discussed and compared with the *Markov Chain Monte Carlo* (MCMC) method [27,30,31]. These methods are ideal for the study of parameter identifiability and allows detection of ambiguous parameter estimates. Despite fundamental differences in theoretical basis, i.e., the PL/PP methods are based on a *frequentist* interpretation of parameter estimation while the MCMC is typical of the *Bayesian* statistics framework, the methods share certain similarities. As shown in Fig. 1, all these methods seek to obtain estimates of the likelihood function $L\left(\theta; y_{[N]}\right)$, or by inclusion of a *prior* $p(\theta)$, the posterior distribution $p\left(\theta|y_{[N]}\right)$. Each method *explores* the parameter space by taking samples θ_k and evaluating them on the same*likelihood/posterior hyper-surface*. However, there are some important differences; the use of *deterministic* vs. *stochastic* exploration of the parameter space, and the use of *projection* in the PL/PP methods vs. *marginalisation* in MCMC to obtain *partial* projections/distributions of selected parameters. An overview of relevant variations of the methods is given in Table 1, together with a short-hand name for each method for future reference. The PL1D/PL2D and PP1D/PP2D are collectively referred to as the PL and PP methods, respectively.

2.2. Parameter estimation and analysis

For simplified models, e.g., thermal network models, the uncertainty in the state transition can be large. Hence, it is convenient to express such models as a grey-box model using a continuous time *stochastic differential equation* (SDE) for the state transitionEq. (1); adopting the notation of [16]:

$$dx_t = f(x_t, u_t, t, \theta)dt + \sigma(u_t, t, \theta)d\omega_t \quad (1)$$

$$y_k = h(x_k, u_k, t_k, \theta) + v_k \quad (2)$$

where $t \in \mathbb{R}$ is the time variable and $x_t \in \mathbb{R}^n_x$ is the continuous time state vector. The first and second terms in Eq. (1) are commonly referred to as the *drift* and *diffusion* term, respectively [16,38]. The *drift* term expresses the deterministic transition of the conditional mean state, while the *diffusion* term expresses the increments of the uncertainty linked to the conditional state covariance. The diffusion term, i.e. the *process noise*, is expressed as the function σ multiplied with the differential of a standard Wiener process ω_t [16,38]. The *measurement equation*, given in Eq. (2) is formulated in discrete time where $v_k \sim \mathcal{N}(0, V)$ is the measurement noise. The continuous time input $u_t \in \mathbb{R}^n_u$ and output $y_t \in \mathbb{R}^n_y$ have the corresponding ordered sequences of discrete time measurements u_k and y_k taken from the system $\mathcal{S}$:

$$y_{[N]} = [y_0, y_1, \ldots, y_N] \quad (3)$$
$$u_{[N]} = [u_0, u_1, \ldots, u_N]$$

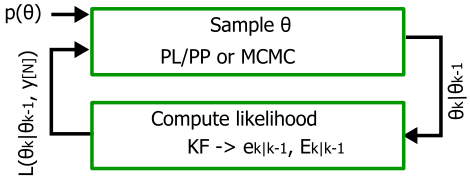

Fig. 1. Both the PL/PP and the MCMC methods explore parameter space on the same likelihood/posterior hyper-surface.

Table 1
Method overview.

Name	Description
PL1D	$L(\theta; y[N])$ *projected* to parameter θ_i
PL2D	$L(\theta; y[N])$ *projected* to plane $\Theta_{i,j}$
PP1D	$p(\theta\|y[N])$ *projected* to parameter θ_i
PP2D	$p(\theta\|y[N])$ *projected* to plane $\Theta_{i,j}$
MCMC	$p(\theta\|y[N])$ *marginalised* to θ_i or $\Theta_{i,j}$

Here, the integer subscripts $k = 0, 1, \ldots, N$ denote the discrete time sampling instants, and the subscript enclosed in $[\cdot]$ is used to indicate an ordered sequence. The estimation of as an optimisation problem, defining the *objective function* $g(\theta)$:

$$\hat{\theta} = \arg\operatorname*{opt}_{\theta} g(\theta; \mathcal{M}, \mathcal{K}, \mathcal{A}) \quad (4)$$
$$s.t.\ \theta \in \Theta$$

Here, $\mathcal{M}$ is a predetermined model structure parametrised by $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^{n_\theta}$ is a set of feasible values for the model parameters that form *inequality constraints* for the optimisation problem. Parameters in θ are sampled from the parameter space Θ by an algorithm $\mathcal{A}$. The experimental conditions $\mathcal{K}$ include the input and output measurements $u_{[N]}$ and $y_{[N]}$ as defined in Eq. (3). In the sequel, the dependency on $\mathcal{M}, \mathcal{K}$ and $\mathcal{A}$ is omitted for simplicity of notation.

A statistically well-founded choice of objective $g(\theta)$ is the *likelihood* function

$$L\left(\theta; y_{[N]}\right) = p\left(y_{[N]} \mid \theta\right) \quad (5)$$

which describes the *joint probability* of observing the measurement sequence $y_{[N]}$ given $\mathcal{M}(\theta)$. An elegant method for evaluating the likelihood function $L\left(\theta; y_{[N]}\right)$ for grey-box SDE models on the form of Eqs. (1) and (2) is presented in the framework named Continuous Time Stochastic Modelling (CTSM) [16]. The CTSM approach is summarised in Section 2.3. An alternative choice for $g(\theta)$ is the *posterior distribution* of the parameters $p\left(\theta|y_{[N]}\right)$, which combines the likelihood, by Bayes' theorem, with a *prior* distribution $p(\theta)$ on the parameters, and with the *evidence* $p\left(y_{[N]}\right)$, a scaling factor that is independent of θ;

$$p\left(\theta|y_{[N]}\right) = \frac{p\left(y_{[N]}|\theta\right)p(\theta)}{p\left(y_{[N]}\right)} \quad (6)$$

Both the likelihood $L(\theta; y[N])$ and the posterior $p(\theta|y[N])$ are *statistical quantities* that relates different values of θ with the *data* $y_{[N]}$, hence representing *density functions*[39]over the parameter space Θ. Observe that, unlike the posterior distribution, the likelihood is *not* a probability distribution over the parameters but takes its *random variable* as the measurements $y_{[N]}$ in the *sample* space, *given* a known parameter θ.

It is interesting to note that the maximisation of the likelihood function is typically associated with a *frequentist* statistics framework, whereas the use of a posterior *distribution* is typical of a *Bayesian* approach. In the frequentist framework, as for the likelihood function, the model parameters are considered constants, while the *data* is the random variable. Hence, the frequentist goal is to estimate some statistic of the *true* parameter θ^*, such as a *confidence interval (CI)* [40,41]. Observe that the confidence level of a CI is *not* a probability statement, as unequivocally stated in [40], since neither the CI nor the true parameter θ^* are considered to be random variables.

In contrast, the Bayesian approach to statistics treats the *parameters* as random variables that are subject to probabilistic treatment, i.e., described by a probability distribution rather than as constants. Typically, the posterior distribution cannot be obtained analytically, and some variation of the *Markov Chain Monte Carlo (MCMC)*method is used instead to *estimate* the posterior distribution of the parameters given the data.

Both the likelihood function and the posterior distribution can be directly optimised to obtain a parameter *point estimate*, respectively denominated the*Maximum Likelihood Estimate (MLE)* $\hat{\theta}_{\text{MLE}}$ and the *Maximum Aposteriori Estimate (MAP)* $\hat{\theta}_{\text{MAP}}$. However, for the purpose of *analysing* the results of the parameter estimation, it is useful to visualise the objective function over the feasible region Θ; either the whole of Θ or some sub-region of particular interest. Since Θ is typically high dimensional, it is necessary to create plots for single parameters, or combinations of two parameters. Since the posterior $p(\theta|y[N])$ is a probability density function (p.d.f.), the posterior for individual parameters or combinations of two parameters can be found by *marginalisation*, i.e., integrating out the remaining parameters. The likelihood function $L(\theta; y[N])$, however, is *not* a p.d.f., and results for individual parameters are therefore obtained by *projections* onto individual parameters or planes of two parameters. These projections can be computed and analysed in the framework of the *Profile Likelihood (PL)* method, typically considered part of the *frequentist statistics* framework [21], in order to diagnose parameter identifiability [10,21,42,5].

If the prior $p(\theta)$ is chosen as *flat*, i.e., a *diffuse* prior is used, $p(\theta) = c$ for $\theta \in \Theta$ and $p(\theta) = 0$ for $\theta \notin \Theta$ where typically $c = 1$, the posterior is *proportional* to the likelihood $p\left(\theta|y_{[N]}\right) \propto p\left(y_{[N]}|\theta\right)$ over the *support* of the prior, i.e., where $p(\theta) \neq 0$, since the *evidence* scaling constant $p\left(y_{[N]}\right)$ is independent of θ. If the prior is *flat* and *unbounded*, i.e., $p(\theta) = c$ for $\theta \in \mathbb{R}^{n_\theta}$, the proportionality $p\left(\theta|y_{[N]}\right) \propto p\left(y_{[N]}|\theta\right)$ holds for all θ. Hence, methods that operate on a *target distribution* $\pi(\theta) \propto p\left(\theta|y_{[N]}\right)$, such as MCMC, can also be used with the likelihood $p\left(y_{[N]}|\theta\right)$ by assuming $p(\theta) = 1$ for $\theta \in \Theta$.

Observe that the use of a *feasible region* $\theta \in \Theta$ is equivalent to selecting a *uniform bounded* prior with a constant value $c = 1$ in the defined space Θ and zero otherwise in $\mathbb{R}^{n_\theta}$. However, the introduction of such a feasible region does not exclude the use of prior distribution $p(\theta)$, since one may well choose $\Theta = \mathbb{R}^{n_\theta}$. If a *non-uniform* prior is used *in addition* to a feasible region Θ, this is equivalent to multiplying the non-uniform prior with a uniform bounded prior $p(\theta \in \Theta) = 1$.

Arguably, by effect of their *omission* in methods operating on the likelihood directly, the use of *flat unbounded* priors is the *default* in the frequentist framework, but it is *non-typical* in Bayesian statistics [42,43]. In practice, particularly in engineering, there is often *some* prior information that could be made use of in the estimation in the form of a prior distribution derived from physical system specifications.

For non-flat priors, many estimation methods based on the likelihood function can be modified to instead optimise on the *posterior* by including the *prior* through Bayes' theorem in Eq. (6). An example of this is the modification of the PL into the PP method presented in [21]. If numerical optimisation is used on the posterior distribution directly, i.e., a prior is included with the likelihood function to form an objective function, the resulting parameter estimate is a MAP point estimate. Indeed, this is supported in the CTSM framework as well [16,22,18].

2.3. Computing the likelihood and the posterior distribution for parameters of grey-box models

Both the MCMC and the PL/PP methods require evaluation of the likelihood function $L\left(\theta; y_{[N]}\right)$, either used directly in PL, or for the evaluation of the posterior distribution $p\left(\theta|y_{[N]}\right)$ in PP and MCMC. The CTSM framework [16,8,17,23] presents a statistically well founded method for computing $L\left(\theta; y_{[N]}\right)$ for grey-box models on the SDE form of Eq. 1.

The likelihood function is defined in Eq. (5). By application of the product rule $P(A \cap B) = P(A \mid B)P(B)$[38], Eq. (5) can be expanded such that [16]:

$$L\left(\theta; y_{[N]}\right) = \left(\prod_{k=1}^{N} p\left(y_k \mid y_{[k-1]}, \theta\right)\right) p(y_0 \mid \theta) \tag{7}$$

In general, evaluating Eq. (7) requires knowing the initial probability density function and successively solving the*Kolmogorov forward equation* [16,38]. However, by assuming a normal distribution for the one-step ahead prediction residuals, a simpler alternative, the multivariate Gaussian distribution, can be used [16]:

$$L\left(\theta; y_{[N]}\right) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\epsilon_{k|k-1}^{T}\mathcal{E}_{k|k-1}^{-1}\epsilon_{k|k-1}\right)}{\sqrt{\det\left(\mathcal{E}_{k|k-1}\right)}\left(\sqrt{2\pi}\right)^{n_y}}\right) p(y_0 \mid \theta) \tag{8}$$

By conditioning on knowing the*initial distribution* $p(y_0|\theta)$, this expression can be iteratively evaluated in a Kalman Filter that estimates the quantities [16,38]:

$$\hat{y}_{k|k-1} = \mathbb{E}\left[y_k \mid y_{[k-1]}, \theta\right] \tag{9}$$

$$\epsilon_{k|k-1} = y_k - \hat{y}_{k|k-1} \tag{10}$$

$$\mathcal{E}_{k|k-1} = \mathbb{E}\left[\epsilon_k \epsilon_k^T\right] \tag{11}$$

where $\hat{y}_{k|k-1}$ is the*predicted output* at time k given the measurements up to and including time $k-1$, i.e., the *one-step-ahead* prediction. The choice of KF implementation depends on the type of state transition model; linear or non-linear, and in the latter case, on the model being differentiable such that the model can be linearised for propagation of the covariance [44].

The assumption of normally distributed residuals can be verified by statistical testing [16,22,17,13]. One possible method is the use of a *cumulated periodogram (CP)*, which by use of plotting indicates if the resulting residuals are reasonably approximated by a normal distribution [16,22,17]. Another, *numerical*, alternative is the use of the *Kolmogorov-Smirnov (KS)*test criterion [13]. The KS criterion can also be used in combination with the CP diagram to compute confidence bounds for the normality assumption on the CP diagram [17]. Other alternatives for normality testing include counting *zero-crossings*, the *auto-correlation function (ACF)* [13], the *inverse* ACF or the *partial* ACF [17].

By taking the negative logarithm, and eliminating the factor $\frac{1}{2}$, the result $\ell_{\mathrm{L}}(\theta) = -2\ln L\left(\theta; y_{[N]}\right)$, where dependency on $y_{[N]}$ is omitted in the sequel for notation simplicity, is obtained as

$$\ell_{\mathrm{L}}(\theta) = \sum_{k=1}^{N} \epsilon_{k1k-1}^{T}\mathcal{E}_{k|k-1}^{-1}\epsilon_{k|k-1} + \ln\left(\det\left(\mathcal{E}_{k|k-1}\right)\right) \tag{12}$$

If instead the *posterior* distribution $p\left(\theta|y_{[N]}\right) \propto L\left(\theta; y_{[N]}\right)p(\theta)$ is chosen, after eliminating the scaling by evidence $p\left(y_{[N]}\right)$ and applying the same transformation as above, $\ell_{\mathrm{P}}(\theta)$ is obtained as:

$$\ell_{\mathrm{P}}(\theta) = \ell_{\mathrm{L}}(\theta) - 2\ln p(\theta) \tag{13}$$

Hence, in log space, the application of a prior $p(\theta)$ is implemented by simply subtracting a value from $\ell_{\mathrm{L}}(\theta)$ that depends *only* on the parameter θ. It is interesting to observe that the use of independent normal prior distributions $\mathcal{N}\left(\theta_{\mathrm{p,i}}, \sigma_{p,i}^2\right)$ for each parameter in $\ell_{\mathrm{p}}(\theta)$ is similar to L^2-norm Tikhonov regularisation [45,46], which indicates that application of non-flat priors can be useful for improving *the generalisation* capability of a calibrated model.

2.4. The stochastic discrete time linear model

For a linear time invariant (LTI) model, which is the form typically used for thermal network models, Eqs. (1) and (2) can be written on discrete time form as [38]:

$$\begin{aligned} x_k &= \tilde{A}x_{k-1} + \tilde{B}u_k + w_k \\ y_k &= \tilde{C}x_k + v_k \end{aligned} \tag{14}$$

where $w_k \sim \mathcal{N}(0, \mathcal{W})$ is the process noise (model error), $v_k \sim \mathcal{N}(0, \mathcal{V})$ is the measurement noise and the discrete time model matrices $\tilde{A} = \exp(\Delta t \cdot A)$ and $\tilde{B} = A^{-1}\left(\tilde{A} - I\right)B$ are computed from the standard linear continuous time model matrices A and B [47,45]. Observe that the three model matrices $\tilde{A}, \tilde{B}$ and $\tilde{C}$, and also the noise covariances $\mathcal{W}$ and $\mathcal{V}$ are typically functions of θ. For the noise covariances $\mathcal{W}$ and $\mathcal{V}$, the square root of the diagonal terms are included in θ, while the off-diagonal terms are assumed zero. This assumption is clearly reasonable for the measurement noise, but also commonly used for the process noise covariance [32]. A further extension on the presented work could be to include the off-diagonal terms of $\mathcal{W}$ in θ as well.

By using the SDE framework outlined in Section 2.3, the noise parameters in $\mathcal{W}$ and $\mathcal{V}$ influence $L\left(\theta; y_{[N]}\right)$ through the computed Kalman gain. In the limit case of zero measurement noise $\mathcal{V} \equiv 0$, the innovation covariance in the KF $\mathcal{E}_{k|k-1} = \tilde{C}X_{k|k-1}\tilde{C}^T$ and the standard equations for the linear Kalman Filter [48] give the Kalman gain

$$K_k = X_{k|k-1}\tilde{C}^T\left(\tilde{C}X_{k|k-1}\tilde{C}^T\right)^{-1} = \tilde{C}^{-1} \tag{15}$$

The aposteriori updated state is

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \tilde{C}^{-1}\left(y_k - \tilde{C}\hat{x}_{k|k-1}\right) = \tilde{C}^{-1}y_k \tag{16}$$

and the one-step ahead predicted output is

$$\hat{y}_{k|k-1} = \tilde{C}\left(\tilde{A}\tilde{C}^{-1}y_{k-1} + \tilde{B}u_k\right) + \epsilon_{k|k-1} \tag{17}$$

Hence, the model in the KF is*treated* as a first order autoregressive model in this limit case. However, the model structure and parametrisation are *still* the same grey-box TN structure and *not* the general black-box structure used in typical *Auto Regressive model with Exogenous input* (ARX) models. Since $X_{k|k} = \left(I - \tilde{C}^{-1}\tilde{C}\right)X_{k|k-1} = 0$, the state estimate covariance $X_{k|k-1} = \mathcal{W}$ and Eq. (8) with $\mathcal{E}_{k|k-1} = \tilde{C}\mathcal{W}\tilde{C}^T$ gives the *weighted least squares prediction error* parameter estimate.

In the limit case of $\mathcal{W} \equiv 0$, indicating a *deterministic* model with no diffusion term, it can be shown that the aposteriori state covariance $X_{k|k} \leqslant \tilde{A}^k X_0\left(\tilde{A}^T\right)^k$ [49,38] which will approach zero for a well-behaved stable system. If the initial state is also *deterministic*, $X_0 \equiv 0$, the state trajectory in the KF is *independent* of the measurements y_k, since $X_{k|k} = 0 \rightarrow X_{k|k-1} = 0$ and therefore

$$K_k = X_{k|k-1}\tilde{C}^T\mathcal{E}_{k|k-1}^{-1} = 0 \rightarrow \hat{x}_{k|k} = \hat{x}_{k|k-1} \tag{18}$$

Hence, Eq. (8) with $\mathcal{E}_{k|k-1} = \mathcal{V}$ gives the *weighted least squares* estimate for a *shooting/ballistic*, i.e., *deterministic*, state trajectory [38].

Both these limit cases are intuitively satisfactory and consistent with the common sense intuition of the KF. Given perfect measure-

ments with $\mathcal{V} \equiv 0$, it is natural to rely exclusively on the measurement at the previous time-step at the expense of the previous estimates. In the case of $\mathcal{W} \equiv 0$, the perfect model *predictions* are trusted and the *data* ignored in the state propagation, with data only used to compute the error.

Since the SDE grey-box framework includes both of these limiting cases, it may arguably be considered a general framework, i.e., an *intermediate* between the purely autoregressive one-step-ahead prediction error and the deterministic output error, depending on the noise parameters. If both noise covariances $\mathcal{W}$ and $\mathcal{V}$ are non-zero, and correctly estimated or known apriori, the Kalman Filter gives the optimal estimate of the state.

Arguably, the limit case of $\mathcal{V} \equiv 0$ results in an LS parameter estimation that is similar to typical black-box methodology, while the limit case $\mathcal{W} \equiv 0$ simulation error is more typical of a white-box modelling approach. Hence, the balance between these two limit cases through the Kalman gain can arguably be considered a mathematical expression for the intermediacy of grey-box models, between the white- and black-box approaches.

2.5. Identifiability of parameters

Since the model structure $\mathcal{M}$ is designed to be a *representation* of a system $\mathcal{S}$, it is often assumed that $\mathcal{S} \in \mathcal{M}(\Theta)$ and that consequently there exists a true parameter vector θ^* such that $\mathcal{M}(\theta^*) = \mathcal{S}$. However, this is rarely the case outside of simulation experiments, since the model structure $\mathcal{M}$ is only an *approximation* of $\mathcal{S}$. In the case of thermal network models based on a *naive* physical approximation of $\mathcal{S}$, the similarity of $\mathcal{M}$ to $\mathcal{S}$ is especially questionable. The estimate $\hat{\theta}$ depends on several factors, such as the amount of dynamic information in $\mathcal{K}$, the choice of objective function $g(\theta)$, and to some extent on the algorithm $\mathcal{A}$. Hence, the subject of *parameter identifiability* is of particular importance for simplified grey-box models, if the estimated parameters $\hat{\theta}$ are themselves of interest.

A model structure $\mathcal{M}$ may be *over-parameterised* such that a subset θ_s of the parameters has no effect on the model predictions $\hat{y}$, either because the model in Eqs. (1) and (2), and therefore also $g(\theta)$, is *free* of certain parameters, or the combined effect of several parameters cancels out. The parameters θ_s, denominated as *structurally* non-identifiable, result in *unbounded* confidence intervals (CI) [10]. Similarly, over-parametrisation may lead to the parameters in θ_s being *inter-dependant*, such that only some functional combination of the parameters are identifiable, resulting in equipotential, possibly bounded, manifolds in the parameter space. Additionally, if the dynamic information content in the data is *insufficient* for estimation of certain parameters, these parameters are*practically* non-identifiable [10]. Based on the definition given in [10,21], parameters are practically non-identifiable when the likelihood is only *somewhat* affected by perturbations of the practically non-identifiable parameters, such that a well-defined optimum exists, but the likelihood is not sufficiently sensitive to produce a *bounded*CI at the desired level of confidence.

The use of CIs as diagnostic criteria for identifiability is a distinctly *frequentist* statistics approach [40]. A formal definition of non-identifiability, based on the Bayesian framework of computing probability distributions of parameters, is given in [25,42]. The subset of *identifiable* parameters is defined such that $\theta = (\theta_i, \theta_s)$. Parameters θ_s are non-identifiable if [25]:

$$p\left(\theta_s|\theta_i, y_{[N]}\right) = p(\theta_s|\theta_i) \Rightarrow \theta_s \perp\!\!\!\perp y_{[N]}|\theta_i \tag{19}$$

That is, no *additional* information is obtained about θ_s from the data $y_{[N]}$ once the identifiable parameters θ_i are known [25]. Hence, the non-identifiable parameters are *conditionally independent* of the data, given the identifiable parameters [25]. Since $p\left(\theta_s|\theta_i, y_{[N]}\right) \propto L\left(\theta_i, \theta_s; y_{[N]}\right) p(\theta_s|\theta_i) p(\theta_i)$, Eq. (19) implies that the likelihood $L\left(\theta_i, \theta_s; y_{[N]}\right)$ is *free*, i.e., unaffected, by θ_s [25,42], which is similar to the description of *structural* identifiability given in [10,21].

As discussed in Section 2.4, the measurement and noise covariance matrices $\mathcal{W}$ and $\mathcal{V}$ are here considered functions of θ. Specifically, the noise covariance matrices are assumed diagonal, with the square root of the non-zero terms included in θ. Identifiability of these parameters is treated in the same way as for the thermal model parameters. For a more thorough analysis of noise model parameter identifiability, see e.g. [35].

2.5.1. Resolving non-identifiability by application of a prior

If both the likelihood and the priors are non-informative for a sub-set of the parameters, there is clearly a problem with the application of *any* parameter estimation method, since there is *no* information from which to estimate the non-identifiable parameters. The solution is to *introduce more information* into the parameter estimation problem, by either redesigning the experiment to obtain more *informative* data and/or new measurements, or by *revising* the model structure to better fit the available data. A third possibility is the addition of a *non-flat* prior distribution, based on prior physical information of the system. Experimental design is particularly challenging for the study of building thermal behaviour since buildings are subject to weather conditions and occupancy demands that are usually beyond experimental control [4]. Hence, the use of priors to resolve non-identifiability is particularly interesting for building thermal modelling.

The local sensitivity of the log posterior distribution in Eq. (13) to perturbations of θ can be estimated by the Hessian:

$$H_P = \nabla^T \nabla \ell_P\left(\theta; y_{[N]}\right)\Big|_{\theta=\hat{\theta}} = H_L - \nabla^T \nabla 2 \ln p(\theta)\Big|_{\theta=\hat{\theta}} \tag{20}$$

where $H_L = \nabla^T \nabla \ell_L\left(\theta; y_{[N]}\right)\Big|_{\theta=\hat{\theta}}$ is the Hessian of the likelihood function[10,16,50]. Hence, if the *likelihood* is insufficiently affected by perturbations of θ in certain directions, as indicated by H_L, the addition of a prior can be seen to introduce another source of sensitivity to perturbations of θ and therefore resolve the non-identifiability. Note that while a prior may resolve non-identifiability and therefore result in unambiguous parameter estimates, it does not necessarily guarantee a physical interpretability of the estimated parameters. Note also that the obtained H_P describes the sensitivity of $\ell_P\left(\theta; y_{[N]}\right)$ which is data dependent [10,16,50].

2.6. Profile likelihood and profile posterior

The PL method [10,5] can be used to estimate uncertainty and diagnose identifiability of the parameters by *projecting* the likelihood function $L\left(\theta; y_{[N]}\right)$ onto each parameter θ_i. The*likelihood profile* $\ell_{PL1D}(\theta_i)$ is defined as the *minimum negative log likelihood* $\ell_L(\theta)$, computed for values of a single parameter θ_i, when the remaining parameters $\theta_{j\neq i}$ are *freely* optimised [10,51]:

$$\ell_{PL1D}(\theta_i) = \min_{\theta_{j\neq i}} \ell_L\left(\theta_{j\neq i}; y_{[N]}, \theta_i\right) \tag{21}$$

Values of θ_i are chosen, either by a brute force discretisation of θ_i or using a gradient decent method, prior to optimising the remaining $\theta_{j\neq i}$ [10]. A likelihood-based CI can be obtained by applying a *threshold* to the likelihood function [10,51]. Let

$$\left\{\theta : \ell_L(\theta) - \ell_L\left(\hat{\theta}\right) < \Delta_\alpha\right\}, \quad \Delta_\alpha = \chi^2(\alpha, n_{df}) \tag{22}$$

where $\hat{\theta}$ is a freely estimated, presumed optimal parameter vector, and the threshold Δ_α is the α percentile of the χ^2-distribution with n_{df} degrees of freedom [52]. By using Eq. (22) to set a threshold on the likelihood profile $\ell_{\text{PL}}(\theta_i)$ of each parameter, it is possible to diagnose parameter identifiability. As discussed in Section 2.5, *structurally* non-identifiable parameters produce *unbounded* CIs, or equivalently, *flat* likelihood profiles [10]. A likelihood-based CI, unlike the Hessian based *asymptotic* CI, is not necessarily symmetric, and can therefore be unbounded in one direction. Hence, a *practically* non-identifiable parameter can be diagnosed if the at least half *unbounded* likelihood profile has a well-defined minimum [10]. Only parameters that produce bounded CIs, and consequently have sufficiently convex likelihood profiles, are identifiable by optimisation of $\ell_{\text{L}}(\theta)$.

The PL method can be extended to project the posterior distribution, rather than the likelihood function, by inclusion of a prior $p(\theta)$ by Bayes' theorem [21]. The PP method is defined, similarly to Eq. (21), by obtaining the *posterior profile* $\ell_{\text{PP1D}}(\theta_i)$ as the *minimum negative log posterior, given in*Eq. (13), for a prescribed value of θ_i when the remaining parameters are freely estimated, i.e.:

$$\ell_{\text{PP1D}}(\theta_i) = \min_{\theta_{j\neq i}} \ell_{\text{P}}\left(\theta_{j\neq i}; y_{[N]}, \theta_i\right) \tag{23}$$

As for the PL method, the posterior profile is obtained for some selected values of θ_i, and subsequently plotting $\ell_{\text{PP1D}}(\theta_i)$. Observe that by replacing the log *likelihood* ℓ_{L} by the log *posterior* ℓ_{P}, the obtained profile is offset by the log of the *prior*, $-2\ln p(\theta_i)$. Finally, observe that the PL method can be considered as a *special case* of the PP method, with the prior $p(\theta) = 1 \rightarrow -2\ln p(\theta) = 0$ for all $\theta \in \mathbb{R}^{n_\theta}$.

2.6.1. Profiling in two parameter dimensions

The typical implementation of the PL/PP method [10,21,5] projects the likelihood/posterior of the n_θ dimensional space Θ onto the single parameter θ_i. These projections are known to *overestimate* the width of the obtained profiles if there are *interdependent* parameters. Hence it is of interest to project the likelihood/posterior in a way that visualises potential parameter interactions. A possible modification of the PL method is then to hold out *two* parameters rather than one, hence the PL2D method obtains [44,45];

$$\ell_{\text{PL2D}}(\theta_i, \theta_j) = \min_{\theta_{k\neq i,j}} \ell_{\text{L}}\left(\theta_{k\neq i,j}; y_{[N]}, \theta_i, \theta_j\right) \tag{24}$$

PL2D projects the log likelihood onto the plane $\Theta_{i,j} = (\theta_i, \theta_j)$ s.t. $\theta_i, \theta_j \in \Theta$. The resulting two-dimensional profiles can be analysed similarly to the one-dimensional profiles [10], using the definition in Eq. (22). The profiles are computed for all combinations of parameters, i.e., by projecting the objective function to all possible planes $\Theta_{i,j}$. Since $\ell_{\text{L}}(\theta)$ is typically similar for neighbouring θ, previous PL2D estimates can be used as a warm-start for new points in $\Theta_{i,j}$ to improve computational efficiency [20]. A*confidence region* in the $\Theta_{i,j}$ plane is obtained by applying the Δ_α threshold from Eq. (22). Observe that since the optimal estimate $\hat{\theta}$ has n_θ free parameters while the PL2D estimate has $n_\theta - 2$, this gives $n_{\text{df}} = 2$ for the computation of Δ_α from the χ^2-distribution in Eq. (22). Based on these two-dimensional profiles, and the computed confidence regions, parameters are considered identifiable if their corresponding confidence regions are bounded in all directions. If the region contains an unbounded equipotential *valley* in the log likelihood space, the parameter is considered structurally non-identifiable. If the profile has a well-defined minima, but is unbounded in one direction, i.e., the log likelihood is below the Δ_α threshold, this indicates a practically non-identifiable parameter [10]. Subsequently, the size and shape of a *bounded* region estimates the *accuracy* with which the parameters can be estimated.

The free estimate $\hat{\theta}$ may with advantage be chosen as the minimum $\ell_{\text{PL2D}}(\theta_i, \theta_j)$ obtained from *all* profiles, since such a search *approximates*, subject to the limitations imposed by discretisation in the brute force exploration, a *free* optimisation of *all* parameters, using the already computed ℓ_{PL2D} results. Since the PL2D profiles cover the entire parameter space Θ, this procedure is less affected by local minima than a direct numerical optimisation.

The PL2D method may also be modified to project the posterior rather than the likelihood, thus the PP2D method projects:

$$\ell_{\text{PP2D}}(\theta_i, \theta_j) = \min_{\theta_{k\neq i,j}} \ell_{\text{P}}\left(\theta_{k\neq i,j}; y_{[N]}, \theta_i, \theta_j\right) \tag{25}$$

This modification is analogous to the extension of the PL1D method into the PP1D method.

2.7. MCMC

The *projection* methods PL1D/PL2D, based on the interpretation of CIs, are typically considered part of a *frequentist* approach to parameter estimation [21]. In the *Bayesian* framework, the goal is to infer a *probability distribution* for the parameter θ, now considered a random variable. Given that the posterior distribution is often not analytically obtainable, the*Markov Chain Monte Carlo (MCMC)* method is instead used to compute an *estimate* of the posterior. Unlike regular Monte Carlo (MC) methods, MCMC draws samples of θ, such that each sample depends on the *previous* sample, by defining a *transition probability* $p(\theta_k|\theta_{k-1})$. If the transition probability is chosen to fulfil the detailed balance equation

$$\pi(\theta_{k-1})p(\theta_k|\theta_{k-1}) = \pi(\theta_k)p(\theta_{k-1}|\theta_k) \tag{26}$$

the generated samples will be drawn *proportional* to the *target distribution* $\pi(\theta) \propto p\left(\theta|y_{[N]}\right)$. Hence, the posterior and its parameters, e.g., mean and covariance, can be approximated by computing the *empirical* distribution as a histogram over the sequence of samples $\theta_{[K]}$. In this work, the MCMC method of choice is the basic *Metropolis* algorithm [30,31] using a normal isotropic *proposal distribution* $\theta_k^c \sim q(\theta_k|\theta_{k-1}) = \mathcal{N}(\theta_{k-1}, \Sigma_q)$ where θ_k^c is a *candidate* for the next step θ_k in the Markov Chain, and Σ_q is the covariance of the proposal distribution, centred on the current step θ_{k-1} [28,27,29]. The work of Hastings [31,53,28], a generalisation of the work of Metropolis [30,28], shows that if the proposal distribution $q(\theta_k|\theta_{k-1})$ is chosen such that it ensures every possible value of θ will *eventually* be visited, and this is combined with an *acceptance probability* test of the generated proposal, the resulting *transition probability* $p(\theta_k|\theta_{k-1})$, constituted of the combined proposal-acceptance scheme, fulfils the requirement of Eq. (26). The acceptance criterion using a normal proposal distribution is defined from the probability ratio:

$$\alpha = \frac{\pi(\theta_k^c)}{\pi(\theta_{k-1})} = \exp\left(0.5\left(\ell_{\text{P}}(\theta_{k-1}) - \ell_{\text{P}}(\theta_k^c)\right)\right) \tag{27}$$

The next step in the Markov Chain is then chosen as θ_k^c with probability $p_{\text{a}} = \min(1, \alpha)$. Observe that α is greater than 1 if the proposal constitutes an *improvement*, in which case the proposal will be accepted with probability 1 [27,28].

2.7.1. Posterior predictive distribution

An advantage of the Bayesian parameter estimation framework, and of the MCMC method, is that the representation of parameter uncertainty, expressed in MCMC as the empirical distribution of the samples $\theta_{[K]}$, enables better estimation of the model's *prediction* uncertainty. The*posterior predictive distribution* can be inferred

from a set of simulated state/output trajectories obtained by Monte Carlo (MC) simulation of the model in Eq. (14).

Note that by using the covariance propagation equations of an LTI system [48], it is possible to compute the uncertainty of the predicted state and output trajectory for a *single* parameter estimate. However, the use of the MCMC sampled set $\theta_{[K]}$ allows accounting for uncertainty in the *parameters*. Additionally, the MC simulation method is not restricted to linear or time invariant systems. For a test dataset of length N, assume x_0 is known with covariance $X_0^{(i)}$ and let $\hat{x}_{0|0} \sim \mathcal{N}(x_0, X_0)$. Then, for each time $k \in [1, N]$ compute

$$\begin{aligned} \hat{x}_{k|0}^{(i)} &= \widetilde{A}\hat{x}_{k-1|0}^{(i)} + \widetilde{B}u_k + w_k \\ \hat{y}_{k|0}^{(i)} &= \widetilde{C}\hat{x}_{k|0} + v_k \end{aligned} \tag{28}$$

where $\hat{x}_{k|0}^{(i)}$ and $\hat{y}_{k|0}^{(i)}$ are the *estimated* future state and output at time k, given only measurement information at time 0, computed using the i-th accepted parameter proposal in $\theta_{[K]}$. The process noise $w_k \sim \mathcal{N}(0, \mathcal{W})$ and the measurement noise $v_k \sim \mathcal{N}(0, \mathcal{V})$ are *drawn* independently at each time-step, for each i-th trajectory, using a random number generator (RNG). The model matrices $\widetilde{A}, \widetilde{B}$ and $\widetilde{C}$, and the covariance matrices $X_0, \mathcal{W}$ and $\mathcal{V}$, are all potentially functions of the i-th sample in $\theta_{[K]}$, hence potentially different for each trajectory. Over these K trajectories, the *distribution* of the *predicted output* for the test set is computed, for each time-step k, as a histogram over the set of estimated outputs $\hat{y}_k^{(i)}, i \in 1, 2, \ldots, K$. A similar approach is used in [32].

2.8. Comparing MCMC and profiling methods

2.8.1. Exploration by drawing samples

The projection based PL/PP methods explore Θ by selecting samples of θ *deterministically*. If a brute force method is used, where the parameter θ_i or the plane $\Theta_{i,j}$ is discretised with a prescribed resolution, the sampled values for each computed profile are completely determined apriori. If a hill-climbing method is used, the next sample is also determined deterministically by evaluating the gradient of the current sample. In contrast, the MCMC method explores the parameter space Θ *stochasticity*, using randomisation to select the next sample, such that each new sample is drawn proportionally to the target distribution $\pi(\theta)$ [29,27,28]. Hence, assuming proper mixing of the chains, the majority of the samples will be drawn from the regions of high posterior density. These are, naturally, the regions of most interest for inference about the parameters [29,27,28]. Subsequently, again assuming proper mixing of the chains, the majority of the computation time will be spent analysing the most interesting regions in Θ.

In contrast, the deterministic brute force sampling of the PL/PP methods explore the parameter space Θ *exhaustive*ly within the prescribed discretisation, which is significantly more time-consuming. The advantage of such exhaustive searches is that they are guaranteed to obtain the global optimum, within the precision allowed by the discretisation of Θ. Additionally, deterministic exploration is unaffected by the flat manifolds caused by non-identifiable parameters, whereas the stochastic exploration of MCMC in such conditions can result in convergence failure for chains of finite length [21]. Observe that the MCMC methods with appropriately selected proposal distributions are also theoretically guaranteed to obtain the global optima for infinite chain lengths [29,27,28]. In practice however, MCMC samples Θ sufficiently for parameter inference even with reasonably short chain lengths.

Since the profiling methods explore the posterior by projections onto individual parameters, or planes of two parameters, the method must be executed repeatedly for each parameter or combination of parameters of interest. This further exacerbates the computational burden. For the one-dimensional projection methods PL1D and PP1D, computation time is linear in the number of parameters n_θ and usually comparable to MCMC. For the PL2D/PP2D methods, however, the computation time is exponential in n_θ, thus, even moderately large numbers of parameters may lead to infeasible computation times.

2.8.2. Projection and marginalisation

Since the MCMC method draws samples in proportion to the target distribution $\pi(\theta) \propto p\left(\theta|y_{[N]}\right)$, the posterior distribution $p\left(\theta|y_{[N]}\right)$, or its hyper-parameters, can be estimated directly on the set of samples $\theta_{[K]}$, e.g., by computing a histogram [28,27,29]. In order to plot the results, the posterior is often presented as *marginalised* distributions over one or two parameters. It is common practice to present *marginal distributions* for all possible combinations of parameters and present the results as *corner plots* [27].

In contrast, the PL/PP methods obtain the estimated profiles by *projecting* the likelihood/posterior onto individual parameters or planes of two parameters. The resulting profiles are *similar* to the *marginalised histograms* obtained by MCMC, but with one important difference. The projections are computed using *optimisation* over the remaining parameters, as illustrated in Eq. (23). This procedure returns the *optimal* density for the given θ_i, or given (θ_i, θ_j) pair for PL2D/PP2D. In contrast, the marginalisation used in MCMC computes the *integral* over the remaining parameters. For some distributions, such as the normal distribution, these two quantities are *proportional*. Hence, if the *scale* of the resulting profiles/distributions is ignored, these methods will, for some cases,result in *similar* profiles/distributions, particularly for the high posterior density regions where the stochastic exploration of MCMC gives the most accurate results.

3. Experimental setup

3.1. Model

Fig. 2 shows a thermal network model structure, which was developed to approximate the thermal behaviour of an experimental building, located at Campus Porsgrunn of the University of South-Eastern Norway (USN). The model is partially based on the R4C2 model presented in [7]. The RC circuit consists of five components: the thermal resistance between room air and wall R_b, the building envelope R_w, and the thermal resistance of windows and doors R_g, and the two capacitances C_b and C_w representing the thermal capacitance of the building interior and envelope, respectively. The model has two outputs: the room temperature T_b and the wall surface temperature T_w, and two inputs: the consumed

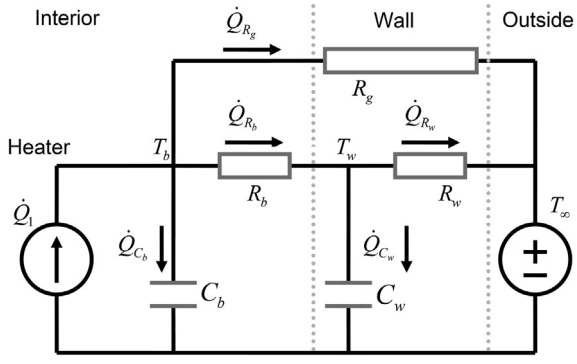

Fig. 2. RC circuit model of the building.

power by an electric heating element $\dot{Q}$ and the outside temperature T_∞. The model can be expressed on the form of Eqs. (14) with state transition matrix A, input matrix B, state vector x and input vector u given as:

$$A = \begin{bmatrix} -\frac{1}{C_b R_b} - \frac{1}{C_b R_g} & \frac{1}{C_b R_b} \\ \frac{1}{C_w R_b} & -\frac{1}{C_w R_b} - \frac{1}{C_w R_w} \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{1}{C_b} & \frac{1}{C_b R_g} \\ 0 & \frac{1}{C_w R_w} \end{bmatrix} \tag{29}$$

$$x_k = \begin{bmatrix} T_b \\ T_w \end{bmatrix}_{t=t_k}, \; u_k = \begin{bmatrix} \dot{Q} \\ T_\infty \end{bmatrix}_{t=t_k}$$

Since all states are observable, the measurement matrix $C = I \rightarrow \hat{y}_k = \hat{x}_k$. The model is LTI, hence a standard KF can be used. The noise covariance matrices $\mathcal{W} = \mathrm{diag}(\mathrm{w}_\mathrm{b}^2, \mathrm{w}_\mathrm{w}^2)$ and $\mathcal{V} = \mathrm{diag}(\mathrm{v}_\mathrm{b}^2, \mathrm{v}_\mathrm{w}^2)$ are also estimated from data, and are assumed diagonal. The parameter vector is then $\theta = [R_g \;\; R_\mathrm{b} \;\; R_\mathrm{w} \;\; C_\mathrm{b} \;\; C_\mathrm{w} \;\; w_\mathrm{b} \;\; w_\mathrm{w} \;\; v_\mathrm{b} \;\; v_\mathrm{w}]$.

A variation of this R3C2 model, is the R2C2 model where the thermal resistance R_g is removed, equivalent to setting $R_g \equiv \infty$ in the R3C2 model.

3.2. Training and test datasets

Fig. 3 shows three *independent* sets of data, collected from the experimental building in February 2018, which consist of three temperature measurements, $T_\mathrm{b}, T_\mathrm{w}$ and T_∞, and one measurement of input electrical power, $\dot{Q}$, supplied to an electric heater. The data has been downsampled to a sampling interval of 30 min. This sample interval was determined experimentally by repeatedly increasing the downsampling ratio and using the PL1 method to test that the downsampled data produced similar results as the higher sample rate original data-set. Note that a sample time of 30 min is arguably reasonable for the main thermal behaviour of a building, but may be excessively long for the heater dynamics and solar gains. However, for this particular data-set, a sample rate of 30 min was found acceptable. The temperatures T_b and T_w are used as reference data for the model outputs, while T_∞ and $\dot{Q}$ are the model inputs. The two training datasets are used for parameter estimation and analysis, while the testset is used only for evaluation of the *posterior predictive distribution*, i.e., to evaluate how well the calibrated model predicts future system behaviour.

3.3. Experiment cases and setup

In the sequel, five different experiment configurations, as listed in Table 2, are analysed and compared.

Case 1 uses the full R3C2 model from Fig. 2 with the priors for all parameters $p(\theta) = 1$ for $\theta \in \mathbb{R}^{n_\theta}$. As the results in Section 4 show, Case 1 results in *non-identifiable* parameters. As discussed in Section 2.5.1, there are several ways to resolve parameter non-identifiability.

Case 2 uses the same model structure, but with the addition of a *prior* on the parameter R_g. The parameter R_g represents the thermal resistance of windows and the door, and can hence be computed by hand. The door in the building has a U-value of $1.2\left[\frac{W}{m^2K}\right]$ and an area of $1.76[m^2]$, while the two windows have U-values $1.3\left[\frac{W}{m^2K}\right]$ and a total area of $1.57[m^2]$. The resulting total UA value is then $4.1\left[\frac{W}{K}\right]$, which gives an estimated thermal resistance $R_g = 0.24$ [34]. The covariance of the prior, i.e., the uncertainty of the estimated mean value 0.24 is *chosen* as 0.01^2. With application of a prior distribution based on physical information of the building, the parameters are shown to be identifiable.

Case 3 instead resolves the non-identifiability by modifying the model structure into the R2C2 model, by removing the parameter R_g from the model and effectively lumping the thermal resistance of windows and the door together with the remaining R_b and R_w. All four parameters of the R2C2 model structure are identifiable, despite using uniform priors. Additionally, Case 3 starts the MCMC chains from the MAP estimate $\hat{\theta}_\mathrm{MAP}$, rather than drawing the initial sample uniformly from the feasible region Θ as is done in Cases 1 and 2, thus negating the need for a burn-in phase in MCMC.

Case 4 uses the same setup as Case 3, except that a random noise component $v_k' \sim \mathcal{N}\left(0, 0.1^2\right)$ is *added* to the data for T_b and T_w prior to analysing the estimated parameters. As the results will show, comparing Cases 3 and 4 reveal some interesting insight into the estimation of *noise covariance parameters* for this model. Case 5 also uses the same setup as Case 3, but now a different dataset, *Training 2*, is used. The other four cases all use *Training 1* for estimation and analysis. For Case 5, however, the Training 2 dataset has slightly more dynamic information content, which, as the results will show, is reflected in the parameter analysis.

For each case, the posterior distribution of the parameters $p\left(\theta|y_{[N]}\right)$ is estimated using the MCMC method. The results are presented as *marginal* distributions, both as one dimensional (1D) for each parameter, and as two dimensional (2D) distributions over two parameters. Additionally, each case is analysed using the profiling methods of Section 2.6 in one and two dimensions in order to obtain *projected profiles* of the log posterior $\ell_\mathrm{P}(\theta)$. Note that the experimental cases use different feasible regions Θ, as evident

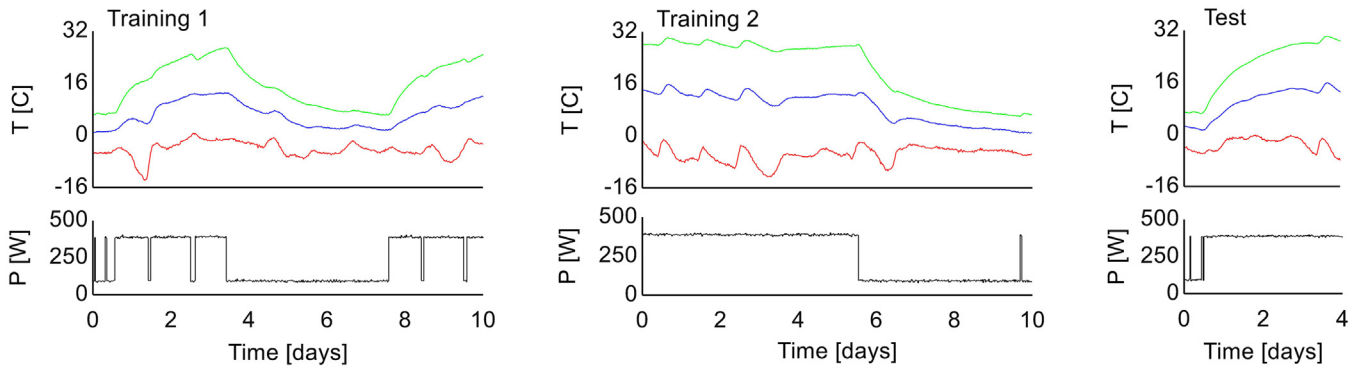

Fig. 3. Training and test datasets, consists of three temperature measurements, T_b (green), T_w (blue) and T_∞ (red), and one measurement of input electrical power, $\dot{Q}$.

Table 2
Configuration for each experiment case.

#	Model	Description
1	R3C2	uniform priors $p(\theta) = 1$
2	R3C2	$p(R_g) = \mathcal{N}\left(0.24, 0.01^2\right)$
3	R2C2	R_g removed, $p(\theta) = 1, \theta_0 = \hat{\theta}_{MAP}$
4	R2C2	Same as 3 + added noise $\sigma = 0.1$
5	R2C2	Same as 3, but using Training 2 dataset

from the result plots in Section 4. Hence, the results for different cases must be compared by taking into account the differences in parameter limits. Note that, for simplicity, the projection method is refereed to in the sequel as PP, since the PL variation can be considered a special case of PP with uniform priors.

3.3.1. Post-processing of results

The marginal posterior distributions from MCMC are *log-transformed*, in the form of Eq. (13), to facilitate comparison with the PP1D/PP2D methods. Additionally, the results are *shifted* in log space such that the minimum of each profile/log distribution is zero, as discussed in Section2.8. Since the goal is to analyse the parameter estimation problem, it is the *shape* in log space and *distribution* over the parameters that are of interest, not the scale or the minimum log posterior value.

The PP1D/PP2D methods, being based in numerical optimisation, naturally respect the bound of the feasible region. In the Bayesian framework, the constraint $\theta \in \Theta$ is equivalent to a prior distribution $p(\theta \in \Theta) = 1$ and $p(\theta \notin \Theta) = 0$, hence in the MCMC implementation any proposed $\theta \notin \Theta$ is automatically rejected.

The results from MCMC and PP2D are presented as *corner plots*, with 2D profiles/marginal posterior distributions for each possible combination of parameters. Additionally, the marginal posterior for each parameter is plotted together with the PP1D profile, for comparison of results.

3.3.2. Tuning

The PP2D method is executed with an experimentally obtained discretisation resolution of 200×200 grid, and a resolution of 400 for PP1D. The MCMC method is applied with 12 chains of length 10^5, except for in the non-identifiable Case 1 which uses chain length 2×10^6 and a thinning factor of 20. For Cases 1 and 2, a fixed burn-in of 10000 samples is used. Cases 3, 4 and 5 initialise the chains at $\hat{\theta}_{MAP}$, hence, no burn-in phase is needed. The proposal distribution is chosen as normal isotropic: $q(\theta_k|\theta_{k-1}) \sim \mathcal{N}\left(\theta_{k-1}, \sigma_q^2\right)$ where $\sigma_q = l \cdot diag(\theta^0)$ and θ^0 is some nominal parameter vector close to the MAP estimate $\hat{\theta}_{MAP}$. The step length $l = 0.01$ has been selected for all cases, such that for Cases 2 to 5 the proposal *acceptance rate* is around 50%. For Case 1, the acceptance rate is found to be around 25, due to the thin elongated valley in the posterior hyper-surface for the non-identifiable case resulting in an increase in rejected proposals.

4. Results and analysis

4.1. Marginal and projected posteriors

Fig. 4 presents the marginal posterior plots from MCMC together with the PP1D and the PP2D projections. Observe that the resulting projections/profiles are *similar* for most parameters. Since the plots are shifted in log space, this similarity indicates *proportionality* in non-log space.

The presence of flat, equipotential regions in the posterior hyper-surface indicates that parameters R_b and R_w are *non-identifiable*. The corresponding PP2D profile in Fig, 4 shows a linear

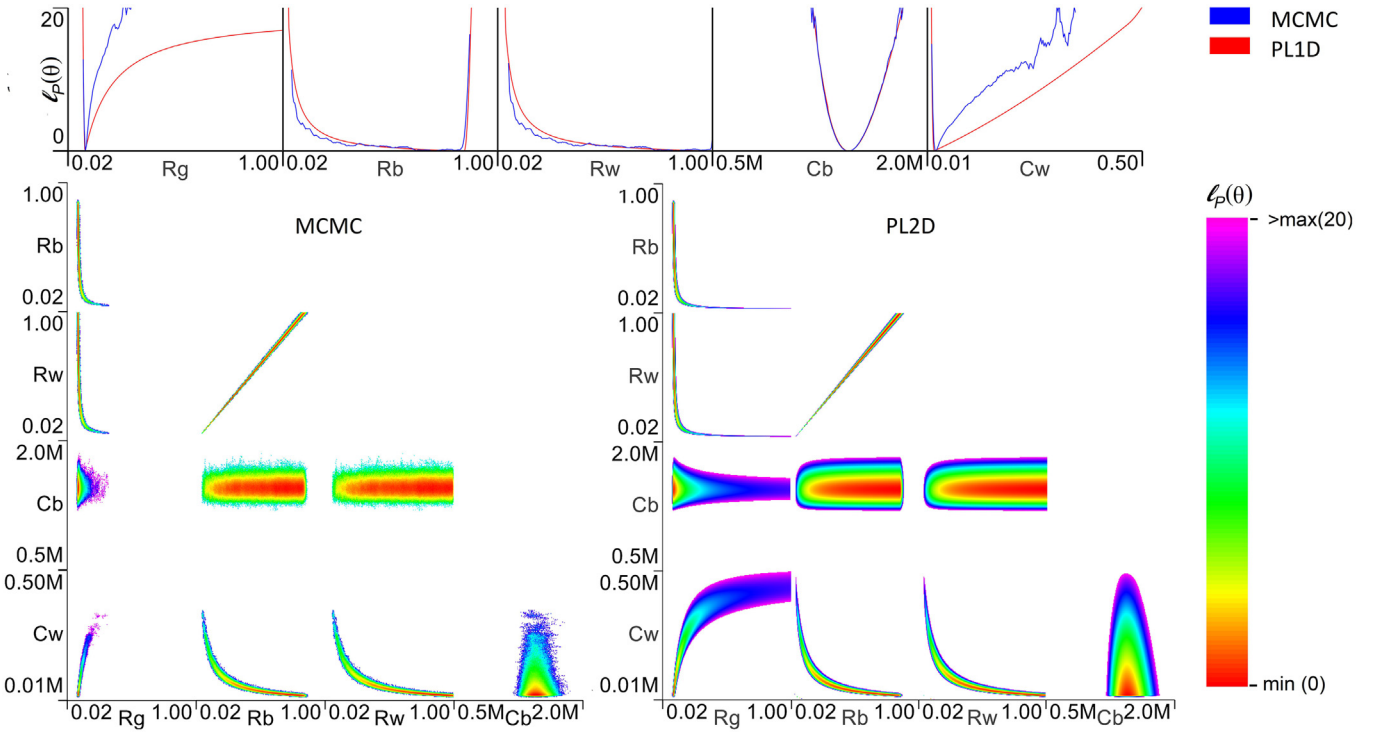

Fig. 4. Comparison of projected profiles from PP1D and PP2D, with the marginalised posteriors obtained from MCMC, for Case 1. The top panel shows the 1D marginals/projections plotted together, where the PL1D projections are plotted in red and the marginalised posterior for each parameter is plotted in blue. The lower left and right panels show corner plots, a set of two-dimensional distributions/projections, one for each possible parameter combination, for the MCMC marginalised posterior and the PL2D projections, respectively.

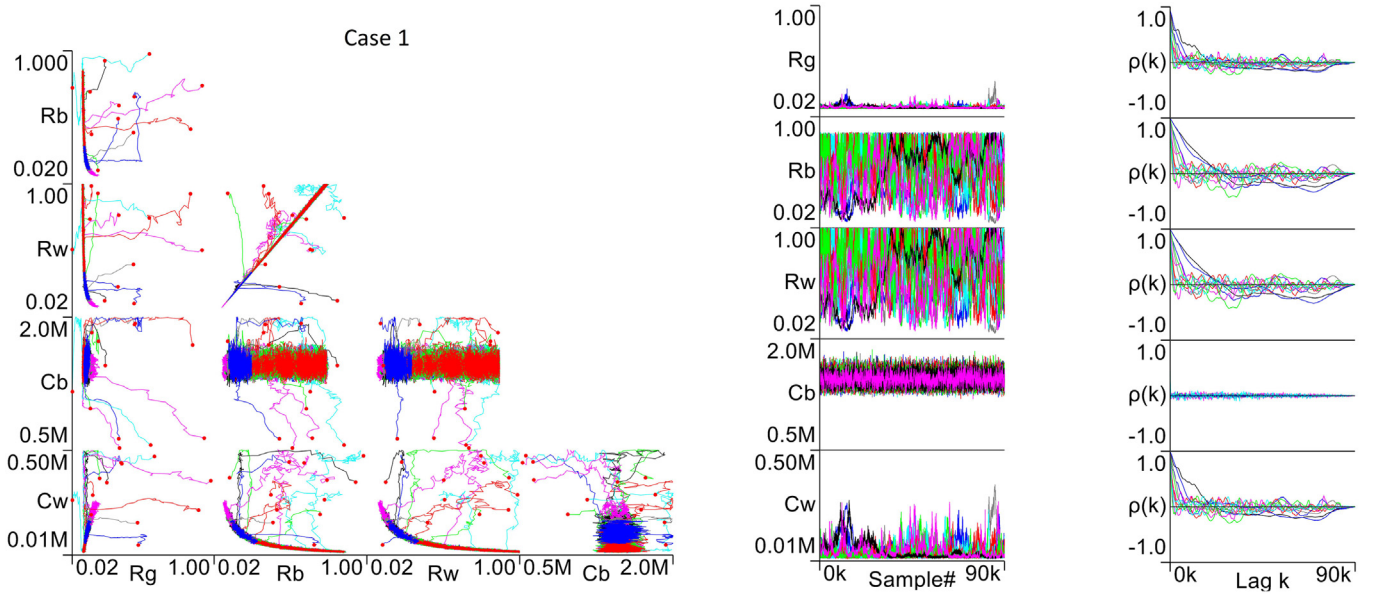

Fig. 5. Diagnostic plots for Case 1: 2D and 1D trace plots are shown in the left and middle panels. The 2D trace plots contain the burn-in phase of 10000 samples, indicating that all chains reach the high-density regions. The right panel shows the ACF plot for each chain after removal of the burn-in phase, which indicates that the MCMC method does not properly converge for Case 1.

inter-dependence between the parameters R_b and R_w, which is indicative of a *structural* problem with the R3C2 model, resulting in non-identifiable parameters because of over-parameterisation.

Next, observe that the marginal posteriors for parameters R_b and R_w show considerable random fluctuation in regions that the PP1D and PP2D methods identify as *flat*. Since MCMC is based on a *stochastic* exploration of the parameter space it is not surprising that the similarity of the results is only *approximate*. The MCMC method draws samples in *proportion* to their posterior density, focusing exploration of Θ on areas of high posterior density, hence the similarity of the results is stronger in these regions.

The differences between projected and marginal posteriors are most pronounced for the parameters R_g and C_w. The marginal posteriors are *not* proportional to the projected posteriors for R_g and C_w, since for these parameters, the optimum obtained by projections is *not* proportional to the integral over remaining parameters. In contrast, the marginal and projected posterior of the parameter C_b are nearly identical in shape.

Since the parameters are subject to the constraint $\theta \in \Theta$, the inter-dependence between R_b and R_w introduces *artefacts* in the profiles, such as the *sharp bend* that occurs in the profile for R_b at ~ 0.9. This phenomenon is caused by the dependant parameter R_w being *actively constrained* < 1.0, hence producing sub-optimal posterior projections for higher values of R_b. The same effect is observed on the MCMC marginal posterior plots, since the bends are caused by the bounds on the parameters and not the analysis method. However, in the Bayesian framework, the constraint $\theta \in \Theta$ can be interpreted as a *prior* on the parameters which *reshapes* the likelihood hyper-surface, consequently resulting in the observed bends in the marginal posteriors of R_b and C_w, as discussed in Section 2.2. The Bayesian interpretation of this phenomenon is arguably more satisfactory than that of *artefacts* induced by active constraints in optimisation.

Because of inter-dependant, and therefore non-identifiable, parameters, the *shape* and *extensiveness* of the posterior hyper-surface become difficult to traverse using the stochastic predict-accept/reject scheme of the Metropolis algorithm. These difficulties are evident by the diagnostic plots given in Fig. 5. For Case 1, the elongated, narrow and flat structure of the posterior hyper-surface for parameters R_b and R_w causes the Metropolis algorithm to sample the posterior somewhat ineffectively, resulting in an average proposal *acceptance rate* of $\sim 25\%$ for the defined proposal distribution $q(\theta_k|\theta_{k-1})$, subsequently with high autocorrelation over the chains. Hence, significantly longer chain lengths were required for Case 1, where $K = 2 \times 10^6$ for all 12 chains, than for the other four cases.

Although the 2D trace plots in Fig. 5 show that all chains quickly reach the high posterior density region for all parameters, from their uniformly drawn starting points $\theta_0 \sim \mathcal{U}(\theta_{min}, \theta_{max})$, the 1D trace plots show that the chains are not reaching equilibrium, except for in the parameter C_b. Since the chains do not converge, the resulting parameter samples $\theta_{[K]}$ are not properly representing the posterior distribution, hence producing less accurate estimates of $p\left(\theta|y_{[N]}\right)$. The autocorrelation function (ACF) plots show significant correlation even at high lag values for all parameters *except* C_b. This is indicative of MCMC chains that are "clumpy" [27,29], as a consequence of failure to converge. Despite the convergence failure, comparison with PP1D and PP2D projected posteriors suggests that the MCMC results are representative of the posterior, although with reduced accuracy.

The resulting marginal distributions and projections of the posterior hyper-surface of Case 2, *reshaped* by the addition of the prior of R_g, i.e., $p(R_g) = \mathcal{N}\left(0.24, 0.01^2\right)$, are presented in Fig. 6. All five parameters are now indicated as identifiable by *bounded* marginal and projected posterior distributions. Despite the different theoretical foundation of the methods discussed in Section 2.6, the marginal and projected posteriors are nearly identical once shifted in log space. The similarity is much stronger than for Case 1, since the challenging equipotential regions in the posterior hyper-surface have been eliminated. As evident from the marginal and projected 2D posterior of parameters R_b and R_w, there is still a strong correlation between them, but there is now a well-defined optimum.

The differences between the use of a *stochastic* rather than *deterministic* exploration of the parameter space is most pronounced in the low posterior density regions. Since the high-density regions are the primary area of interest for these

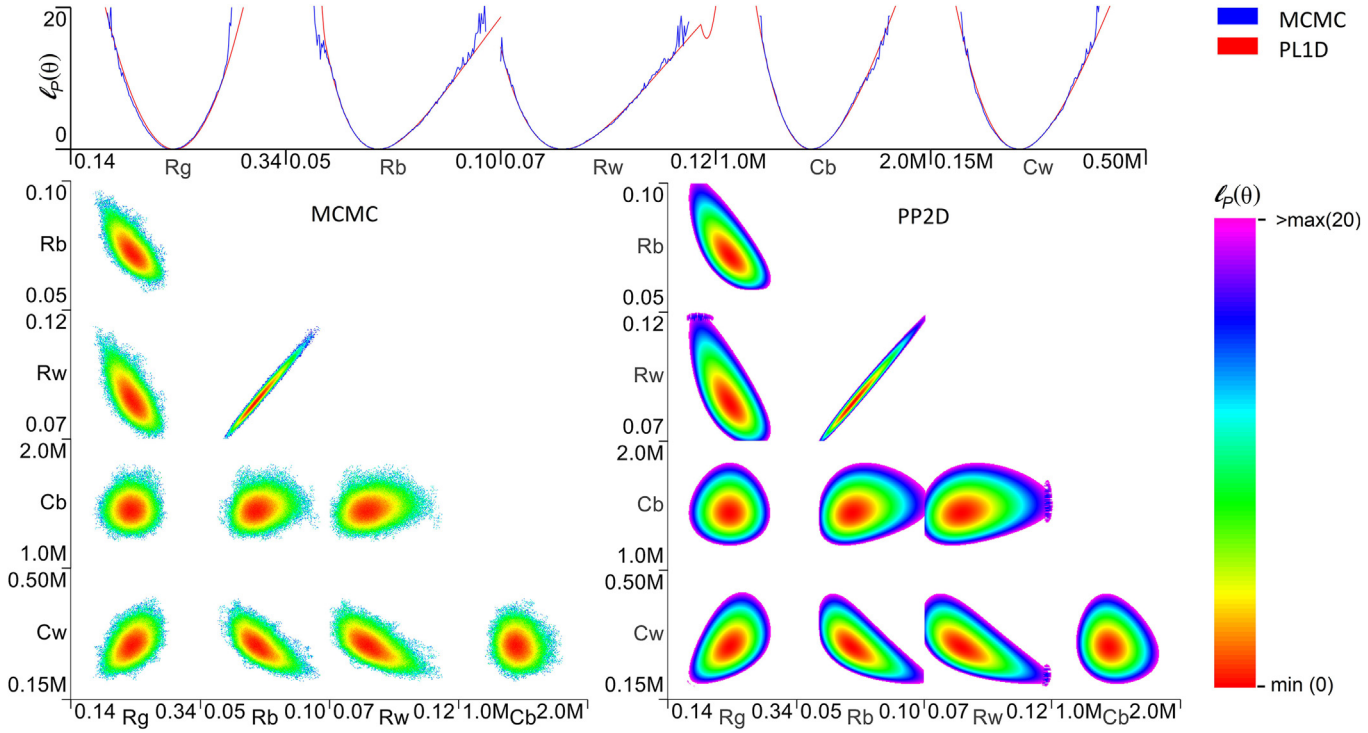

Fig. 6. Marginal and projected posterior for Case 2, with left and right panels showing the MCMC and PP2D results, respectively. A comparison between the 1D marginal distributions (blue) and the PP1D projections (red) is shown in the top panel.

analyses, the somewhat random exploration of the low posterior density regions of MCMC is of no practical consequence. Hence, both methods arguably provide the *same* insights of the parameter space of Case 2. Since the parameters now have well-defined optima, convergence of the MCMC chains occurs well within the predetermined burn-in phase, hence shorter chain lengths where required for Case 2.

The results for Case 3 are presented in Fig. 7. As shown by the marginal and projected posteriors, the posterior hyper-surface has been further reshaped by the removal of R_g. As for Case 2, all

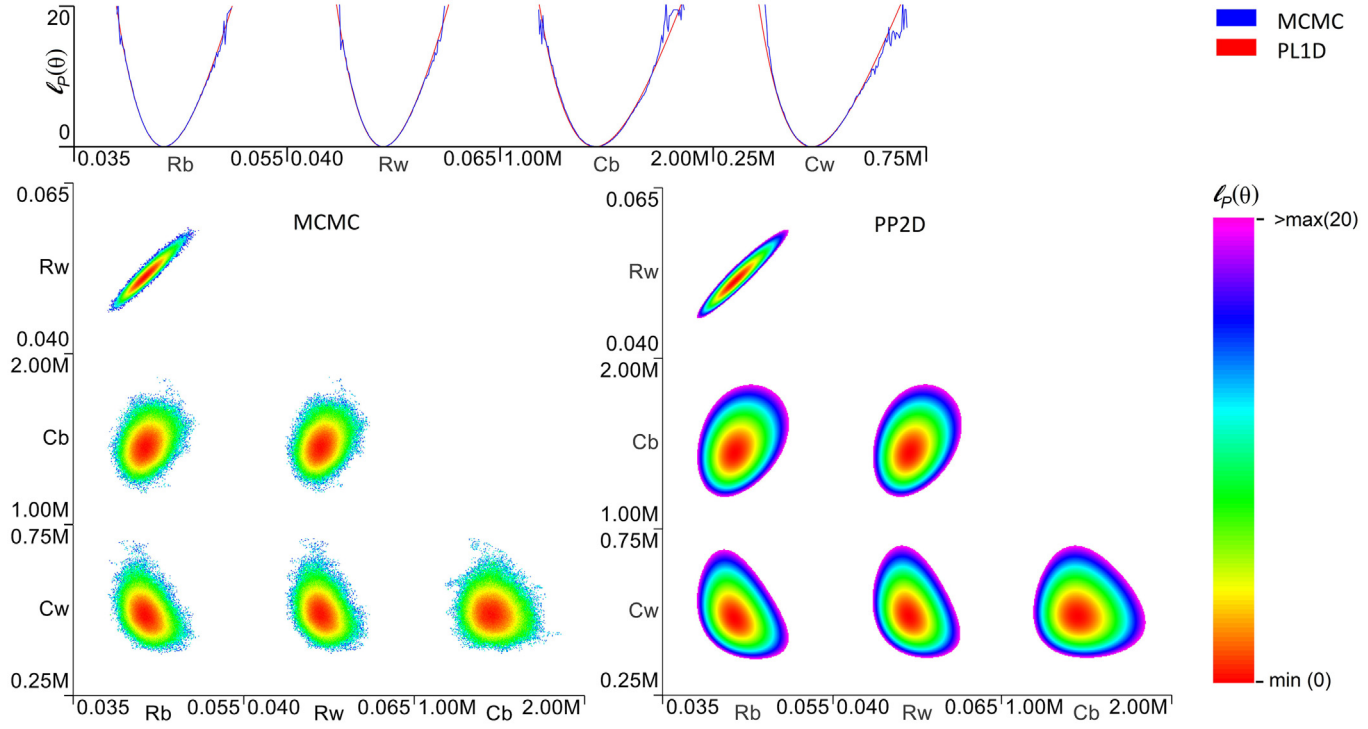

Fig. 7. Marginal and projected posterior for Case 3, with left and right panels showing the MCMC and PP2D results, respectively. A comparison between the 1D marginal distributions (blue) and the PP1D projections (red) is shown in the top panel.

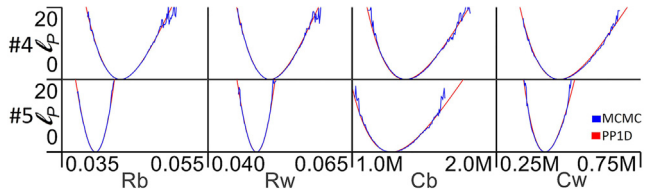

Fig. 8. 1D marginal distributions (blue) and the PP1D projections (red) for Cases 4 (top) and 5 (bottom).

parameters have bounded profiles and are therefore identifiable. However, the *uncertainty*, i.e., the span of the log posterior projections and marginal distributions, is greatly reduced for Case 3. Hence a reduction in the region of interest Θ is required, compared to Case 2, as shown in the ranges of the plots in Fig. 7.

Next, observe that the similarity between the marginal and projected posteriors is stronger for Case 3 compared with Case 2, although there are still some minor differences in the low posterior density regions due to the stochastic exploration of MCMC. These results further confirm that the two methods produce results that are *proportional* for the reshaped hyper-surface of Case 3. Hence, both methods provide the same insight into the parameter estimation problem.

The 1D marginal and projected posterior results for Cases 4 and 5 are shown in Fig. 8. Since the model structure and prior configuration are the same as for Case 3, the structural identifiability and parameter inter-dependency are also the same. The results for Case 4 are nearly identical to Case 3, but Case 5 obtains slightly different MAP estimates and uncertainties, since a different dataset is used for the parameter estimation.

4.1.1. Noise parameters

When calibrating grey-box thermal network models for the purpose of using the estimated parameters to classify building thermal behaviour, naturally, the thermal resistance and capacitance parameters are of primary interest. However, in this paper, the parameters of the noise covariance matrices $\mathcal{W}$ and $\mathcal{V}$ are also estimated from the data. Hence, it is interesting to study the identifiability of the noise parameters; the square root of the diagonal elements of each noise covariance matrix.

The PP1D projected profile and the MCMC marginal 1D posterior for noise parameters w_b, w_w, v_b and v_w for all five cases are presented in Fig. 9. First, observe that the noise parameters for Cases 1, 2 and 3 are nearly identical, despite some of the *thermal* parameters of Case 1 being non-identifiable. Observe also that the projections/marginal distributions are quite similar, indicating that similar information is obtained by both PL/PP and MCMC methods also for the noise parameters.

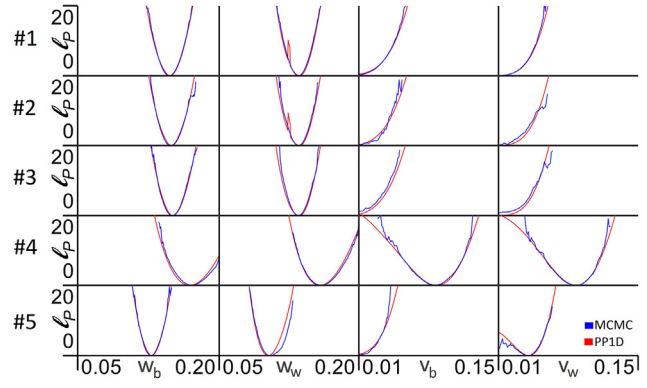

Fig. 9. 1D marginal distributions (blue) and the PP1D projections (red) for the noise parameters w_b, w_w, v_b and v_w for all five cases, presented in increasing order.

Next, observe that the profiles for the measurement noise parameters in $\mathcal{V}$ are *unbounded* towards the minimum of Θ. As discussed in Section 2.4, the noise parameters influence the likelihood *through* the computed *Kalman gain*. Since the Kalman filter estimated state trajectory is optimal when both $\mathcal{W}$ and $\mathcal{V}$ are correctly estimated [48], the values of all four noise parameters are structurally identifiable. However, if the optimal estimate of the measurement noise $\mathcal{V}$ is much smaller than the process noise $\mathcal{W}$, the Kalman gain approaches the inverse of the measurement matrix, i.e. $K \to \tilde{C}^{-1}$. Hence, the updated state depends almost exclusively on the measurement, such that $\hat{x}_{k|k} \approx \tilde{C}^{-1} y_k$. If the model uncertainty is indeed much larger than the measurement uncertainty, relying exclusively on measurements to update the state trajectory is arguably reasonable. However, this results in *practically* non-identifiable measurement noise parameters, since estimating lower values for the elements of $\mathcal{V}$ only drives K slightly closer to $\tilde{C}^{-1}$, and therefore only produces an upper bound on v_b and v_w. This effect can be observed in Fig. 9 for both measurements in Cases 1, 2 and 3, and for measurement T_b in Case 5. For Case 4, with the addition of artificial noise, the measurement noise parameters are both structurally and practically identifiable with well-defined minima and bounded CIs.

4.2. MAP point estimates with uncertainty

Since the posterior hyper-surface for most parameters and experimental cases is known to be asymptotically Gaussian [17], the uncertainty of the MAP estimate for identifiable parameters can be estimated by the Hessian from Eq. (20), i.e., the *covariance* of the MAP estimate is $\Sigma_{\theta_{MAP}} = 2H_P^{-1}$. The $\hat{\theta}_{MAP}$ estimate and estimated standard deviation $\sigma_i = \sqrt{\Sigma_{i,i}}$ are shown, together with normality test results of the $\hat{\theta}_{MAP}$ estimate, in Table 3. The parameter values enclosed in $(\cdot)$ are *ambiguous,* previously diagnosed as non-identifiable, and used only to test the residuals for normality. The corresponding uncertainty estimates are noted as $\times$. The elements denoted as n/a are not relevant due to use of the R2C2 model. The standard deviation σ is normalised over the $\hat{\theta}_{MAP}$ estimate to facilitate comparison of different parameters. The residual for the $\hat{\theta}_{MAP}$ estimate for all five cases passes both the Zero-Crossing (ZC) test with acceptance range $\langle 219, 262 \rangle$and the Kolmogorov-Smirnov (KS) test with threshold (< 0.062),both at confidence $\alpha = 95\%$. The three resistance parameters are given in unit $[K/W]$, the two capacitances in unit $10^6[J/K]$ and the four noise parameters in unit $[K]$.

There are several interesting observations to be made from Table 3. First, comparing Case 2 and 3, lumping R_g into the remaining resistance parameters R_b and R_w results in correspondingly *decreased* estimates of thermal resistance. Next, observe that the MAP estimate of C_b, and the corresponding uncertainty, is approximately the same for all five cases, although slightly lower for Case 5. Note also that the model prediction uncertainty parameters w_b and w_w are nearly identical for the first three cases. These observations indicate at least some correlation of the estimated parameters to the physical properties of the building, which is further discussed in Section 4.4.

Further, observe that although the inclusion of a prior on R_g in Case 2 produced unambiguous MAP estimates, the uncertainty of the remaining estimated parameters is significantly lower in Case 3, where $R_g = \infty$.

Comparing Cases 3 and 4, the uncertainty in the four thermal parameters is not significantly affected by the addition of artificial measurement noise in Case 4. This comparison indicates that a slight increase in measurement noise does not adversely affect

Table 3
MAP parameters with normalised standard deviations computed with the Hessian method, together with normality test results from using Zero-Crossing (ZC) and Kolmogorov-Smirnov (KS) on residuals using the $\hat{\theta}_{\text{MAP}}$ estimate.

#		R_g	R_b	R_w	C_b	C_w	w_b	w_w	v_b	v_w	Output	T_b	T_w
1	$\hat{\theta}_{\text{MAP}}$	(0.101)	(0.515)	(0.607)	1.449	(0.041)	0.148	0.137	(0.000)	(0.004)	*ZC*	247	253
	$\frac{\sigma}{\theta_{\text{MAP}}}$	×	×	×	4.6%	×	3.2%	3.6%	×	×	*KS*	0.035	0.049
2	$\hat{\theta}_{\text{MAP}}$	0.236	0.072	0.084	1.444	0.293	0.149	0.137	(0.003)	(0.002)	*ZC*	243	253
	$\frac{\sigma}{\theta_{\text{MAP}}}$	6.0%	5.7%	5.7%	4.7%	8.8%	3.2%	3.7%	×	×	*KS*	0.039	0.050
3	$\hat{\theta}_{\text{MAP}}$	n/a	0.043	0.051	1.446	0.481	0.151	0.136	(0.010)	(0.010)	*ZC*	243	253
	$\frac{\sigma}{\theta_{\text{MAP}}}$	n/a	2.6%	2.7%	4.9%	7.1%	3.2%	4.9%	×	×	*KS*	0.038	0.050
4	$\hat{\theta}_{\text{MAP}}$	n/a	0.043	0.051	1.369	0.486	0.169	0.158	0.088	0.088	*ZC*	246	259
	$\frac{\sigma}{\theta_{\text{MAP}}}$	n/a	2.7%	2.8%	5.3%	8.0%	5.7%	5.5%	11.5%	10.7%	*KS*	0.030	0.027
5	$\hat{\theta}_{\text{MAP}}$	n/a	0.040	0.048	1.270	0.419	0.128	0.103	(0.002)	0.041	*ZC*	247	259
	$\frac{\sigma}{\theta_{\text{MAP}}}$	n/a	1.4%	1.5%	6.3%	4.5%	3.2%	5.2%	×	18.0%	*KS*	0.052	0.041

the parameter estimation uncertainty. However, the use of a different dataset in Case 5 significantly *reduces* the uncertainty of the thermal parameters. The important factor determining the uncertainty of the estimated parameters is the *dynamic information* content related to the system behaviour contained in the data, assuming a reasonable signal to noise ratio. Finally, observe that the MAP estimates of the four thermal parameters are in reasonable agreement for Cases 3, 4 and 5, which indicates that the estimated parameters are consistent irrespective of the data-set used for calibration, at least to some degree considering the datasets where recorded consecutively.

4.3. Posterior predictive distribution for the Test dataset

Fig. 10 shows the posterior predictive distributions discussed in Section 2.7.1, for each experimental case, computed by *Monte Carlo (MC) simulation* of the model in Eq. (14) for a *thinned* subset of the parameter samples in $\theta_{|K|}$. To reduce computation time, a thinning factor of 100 is used for this computation. The plots are created by repeatedly simulating the test-set ballistically, with randomly generated measurement and process noise v_k and w_k, thus creating one simulated trajectory for each $\theta \in \theta_{|K|}$. The lower right plot shows the standard deviation σ_k of at each time step over the set of K ballistic simulations such that

$$\sigma_k^2 = \mathbb{V}\left[\hat{y}_{k|0}\right] = \frac{1}{K-1}\sum_{i=1}^{K}\left(\hat{y}_{k|0}^{(i)} - \bar{y}_{k|0}\right)^2$$

where $\bar{y}_{k|0} = \mathbb{E}(\hat{y}_{k|0}) = \frac{1}{K}\sum_{i=1}^{K}\hat{y}_{k|0}^{(i)}$. Note that the test-set *measurements* of future system inputs u_k are usedto compute the posterior predictive distributions in order to separate the uncertainties of the model predictions with those introduced by using more realistic *predicted* system inputs.

First, observe from Fig. 10 that Cases 1, 2 and 3 produce *similar* prediction results, despite the differences in model structure and parameter posterior distributions, and from the lower right panel showing the standard deviation (SD) that the empirical SD is nearly identical for the first three cases. Comparing cases 4 and 5 to Case 3 shows that the SD of the predictions is *increased* for Case 4 but *decreased* for Case 5. Since Case 4 has artificially added measurement noise, it is expected that the output predictions will have increased uncertainty, due to larger values of the generated measurement noise parameters v_b and v_w. For Case 5, the variance of the output trajectories is reduced, since the estimated parameters in $\theta_{|K|}$ have less variation due to improved dynamic information in the *Training2* dataset. The similarity of the model predictions for each case is further demonstrated by the root mean square error (RMSE) of the MAP predictions shown in Fig. 10. The observed differences

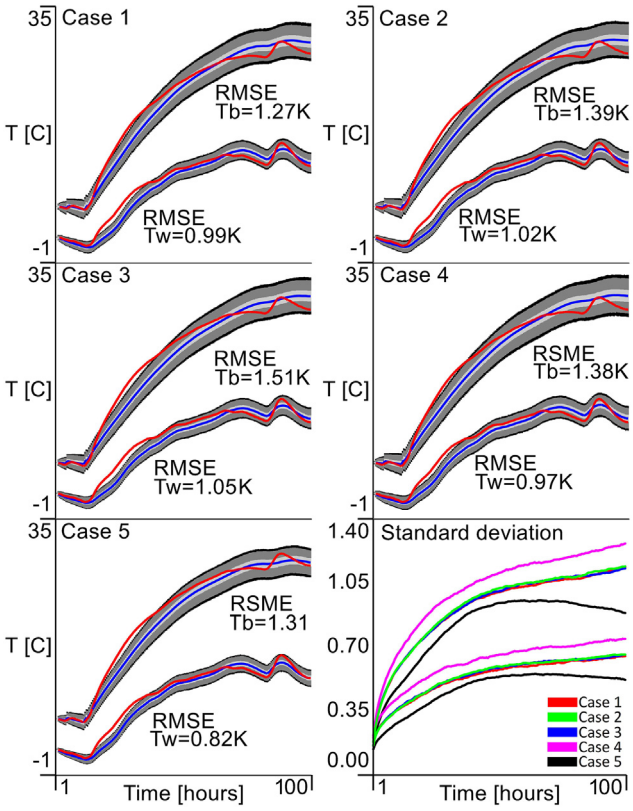

Fig. 10. Prediction posterior from repeated ballistic simulations over $\theta_{|K|}$ of the Test dataset for all five cases. For each case, the temperatures T_b and T_w are plotted as upper/lower temperature, respectively. The plots show the MAP estimate (blue), credibility bands of 50%, 95% and 99% (light to dark) and the reference measurements (red). The lower right plot shows the standard deviation of the outputs, computed over all K trajectories, at each time step.

in predictions over the four-day horizon are likely to be of no practical significance for use in a predictive controller.

The similarity of the posterior predictive distributions of Cases 1, 2 and 3 shows that all three variations of the model are in fact able to *learn* the information in the training set necessary to *predict* the test set. The fact that Case 1 has some non-identifiable parameters with a significant equipotential region in their posterior distributions does *not* prevent the model from successfully predicting the output. Comparing Cases 4 and 5 to Case 3 further indicates that it is the dynamic information contentin the training data that most significantly affects the posterior predictive distributions, as long as the model structure is sufficiently complex to learn the appropriate system behaviour.

These results show that the presented grey-box model may *adequately predict* the system behaviour, even if the parameters are not *unambiguously* identifiable, and that the prediction accuracy largely depends on the information content in the training data. For black-box models, there is usually no assumption of physical interpretability of the model coefficients, hence unambiguous optimal parameter estimates are of no consequence. Methods such as system identification [11] and Artificial Neural Networks (ANN) [54,55] typically produce non-unique system description models whose ability to predict future data, assuming adequate model complexity, depends mostly on the information content in the training data.

By including the stochastic process and measurement noise terms in the model and learning their parameters from data, the model predictions can also reflect these important uncertainties. The computation of a posterior predictive distribution, rather than a single MAP or MLE trajectory, could *facilitate* use of *stochastic* MPC methods [56]. Calculating the *simulation* RMSE to a worst case error of $1.5K$, computed over a reasonably long prediction horizon of four days is likely sufficient for the purposes of model based control [3].

4.4. Physical interpretation of estimated parameters

Given that all five cases show similar predictive capabilities, despite Case 1 having some non-identifiable parameters and Table 3 showing large variation in the model parameters, it is interesting to consider if there are any similarities between the cases that are *not* expressed in the model parameters. Three properties of interest are the eigenvalues λ_1 and λ_2 of A, or rather their negative inverse, i.e., time-constants $T_1 = -\lambda_1^{-1}$ and $T_2 = -\lambda_2^{-1}$, and the total resistance to heat-loss $R_{\text{TOT}} = R_g||(R_b + R_w)$ between the indoor temperature T_b and the outdoor temperature T_∞, where $||$ indicates a *parallel* connection of resistors, i.e., a *harmonic* sum. Note that for Cases 3, 4 and 5, $R_g = \infty \rightarrow R_{\text{TOT}} = R_b + R_w$.

Since the posterior distribution generated by MCMC is represented by a set of samples $\theta_{[K]}$, the quantities T_1, T_2 and R_{TOT} can be computed for each sample in $\theta_{[K]}$ and the marginal posterior distribution for each quantity computed by histogram. The marginal log posterior $\ell_P(\theta) = -2 \ln p\left(\theta | y_{[N]}\right)$ for each quantity is given for all five cases in Fig. 11, together with the MAP estimate and 2.5 / 97.5 credibility percentiles, computed from interpolation on the cumulative empirical distribution.

First, observe that these quantities have bounded profiles with a well-defined optima, also for Case 1. Even though the MCMC method's trace plots in Fig. 5 show a large variation in the model parameters, the time-constants and total resistance are well-defined. When MCMC proposes a new sample θ_k^c, if that sample gives time-constants T_1 and T_2, or a total thermal resistance R_{TOT} that differs *substantially* from the MAP estimates shown in Fig. 11, the resulting log posterior $\ell_P(\theta)$ would produce a very low acceptance probability α.

Next, observe that all three quantities are in reasonable agreement for the first four cases, with the MAP estimate for each case falling within the credibility limits for all the other cases, except the MAP of T_1 for Case 1 falling just below the 2.5% credibility limit of Case 3. The similarity, despite using different model structures, priors and noise on training data, indicates that the time-constants and total resistance are *somewhat* invariant to the experimental setup. The consistency of $R_{\text{TOT}} \sim 0.094$ for Cases 1 to 4 explains why there is a strong correlation between R_b and R_w as illustrated in Section 4.1. Given that R_g is determined by the prior, omitted, or has a well-defined optimum, the values of R_b and R_w must fulfil $R_{\text{TOT}} \sim 0.094$. However, it is difficult to see any *physical* reason for this correlation.The interpretation of the individual R_b and R_w parameters as physical properties of the building is therefore questionable.

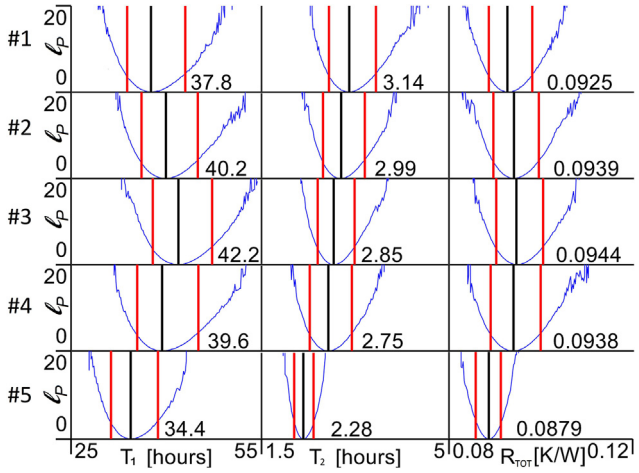

Fig. 11. Log posterior marginal distributions with credibility 2.5 and 97.5 percentiles (red) MAP estimates (black) of the two time-constants, T_1 and T_2, and the total heat-loss resistance R_{TOT} of the building.

Finally, note from the marginal posterior distributions in Fig. 11 that the differences between Case 5 and Cases 1 through 4 are *significant*, e.g. a value of $R_{\text{TOT}} = 0.094$ would give a very low posterior probability based on the distribution of Case 5. The Case 5 MAP estimate for each quantity is outside the 2.5/97.5 percentiles for Cases 1 to 4. While the results of Case 5 *cognitively appear* similar to Cases 1 to 4, the presented results are not similar *enough* to conclude with sufficient statistical *credibility* that the parameters are consistent also when different calibration data is used.

4.5. Comparing MCMC and PP methods

As the results from the five experimental cases have shown, the MCMC and PL/PP methods indeed provide similar results. Even though the theoretical foundation of the methods differ significantly, in particular in the stochastic vs. deterministic exploration of Θ and the use of projection vs. marginalisation to present results, both methods produce estimates of the same posterior parameter distribution. Note, however, the projected and marginalised posterior is *not* always proportional as shown in Case 1. The advantage of the projection-based methods is mainly that they are not affected by flat regions in the likelihood function or posterior distribution. Additionally, the deterministic projections of each prescribed point in $\Theta_{i,j}$ allow an *exhaustive* exploration of the feasible region Θ.The method will therefore obtain both global and local minima, including any equipotential manifolds. The main advantages of MCMC are computational speed and the way that the target distribution is represented as a set of samples $\theta_{[K]}$ that can be used for further analysis, such as computing derived parameters, e.g. time constants or total thermal resistance.

4.5.1. Computation time

Deterministic brute force exploration is, naturally, quite time-consuming. The accuracy at which a global optimum can be found depends on the resolution of the parameter discretisation used in the brute force grid exploration. Hence, the key to successful use of the PL2D/PP2D methods is a reasonable compromise between computation time and resolution. The computational burden is further exacerbated by the need to project the log likelihood or log posterior to all parameter combinations $\Theta_{i,j}$ of interest. In contrast, the MCMC method is specifically designed to explore the most interesting areas, i.e., the areas of Θ with the highest posterior den-

sity. Additionally, since the resulting 2D distributions are computed from histograms by marginalising out the other parameters, there is no need to run the method multiple times. All computation times discussed here are given for the method configurations stated in Section 3.3. All the methods discussed in this paper can easily be parallelised, and can thus take advantage of modern multi-core CPU architectures.

The MCMC method computes in around $\sim 12min$ for Cases 2 through 5.Due to non-identifiable parameters, Case 1 required around 20 times long chains to produce a reasonable approximation of the posterior, thus taking a computation time of $\sim 4.5h$. In contrast, the PP2D method takes around $\sim 6.5h$ to compute all ten projections for Cases 1 and 2. The PP2D method is not affected by the shape of the posterior surface. Case 1 and 2 therefore takes approximately the same time to compute. The six projections in Cases 3, 4 and 5 are computed in around $\sim 2.5h$. The reduction in computation time is due to a lower number of parameters in the last three cases, resulting in fewer projection planes $\Theta_{i,j}$ and also fewer free parameters to optimise for each projected point. The PL1D/PP1D methods, requiring discretisation only of single parameters and only one projection per parameter, are significantly faster at around $\sim 3min$. Note that if only 1D posterior distributions were of interest, the MCMC method could likely have been configured with significantly shorter chains.

4.5.2. Predictive posterior and combined parameter distributions

A distinct advantage of MCMC is the ability to compute posterior predictive distributions for the model output. By using MC simulations over the set $\theta_{[K]}$, a set of K independent state and output trajectories can be computed. The uncertainty of the predictions, given both model uncertainty $w_k \sim \mathcal{N}(0, \mathcal{W})$, measurement uncertainty $v_k \sim \mathcal{N}(0, \mathcal{V})$ and the uncertainty in the parameter estimates as expressed in $\theta_{[K]}$, can be estimated for each time-step over the K trajectories, as discussed in Section 2.7.1.

Another use of the sampled set $\theta_{[K]}$ is the possibility to compute combined parameters, such as the eigenvalues of A or the total resistance to heat-loss R_{TOT} discussed in Section 4.4. Marginal distributions for these combined parameters can then be computed and analysed to provide a more flexible analysis and deeper insight into the model's behaviour.

5. Conclusion

In this paper, both *frequentist* and *Bayesian* frameworks for parameter estimation were used to obtain a detailed analysis of the parameter space of a grey-box thermal network model for a building [43,21]. The Profile Likelihood (PL), the Profile Posterior (PP) and the Markov Chain Monte Carlo (MCMC) methods were used to estimate the *shape* of the posterior distribution for the parameters of a thermal network grey-box model expressed as a stochastic differential equation (SDE) [16].

Five experimental cases were investigated, one of which has non-identifiable parameters. This non-identifiability was shown to be resolved by application of either a *prior distribution* for the parameter R_g, or by the *removal* of R_g from the model, in Case 2 and 3, respectively. Cases 4 and 5 showed how, and under what conditions, the covariance of the process uncertainty w_k and the measurement uncertainty v_k can be estimated from data for the given model. By using the sampled set $\theta_{[K]}$ from MCMC, the eigenvalues, and subsequently the time-constants, of the state transition model A, and also the total thermal resistance R_{TOT}, were shown to have well-defined *bounded* distributions even for the non-identifiable Case 1. The estimates of the time-constants and total thermal resistance were found to be similar for the first four cases, but with significant differences in Case 5, which used a different training dataset.

A distinct advantage of the MCMC method is the ability to use the sampled set of parameters $\theta_{[K]}$ to propagate the uncertainty of the parameter estimates into the model output predictions, by computing the *posterior predictive distribution*. The resulting distributions for all five cases were found to be in reasonable agreement. Hence, all the models are found able to *learn* the necessary *knowledge* about the physical building from the training data necessary to predict the independent test set. This result indicates that while *parameter identifiability* is important for justifying a physical interpretation of the model parameters [5], the presented model's ability to predict system behaviour is not significantly affected by non-identifiable parameters. This result is well-know from the black-box modelling paradigm [11]. Since grey-box models explicitly applies prior physical knowledge of the system to create a model structure, the interpretation of parameters as physical constants of the system is often assumed [5]. The results presented here show that, even if the model correctly predicts the system behaviour, assumptions of physical interpretation of parameters should be supported by an identifiability analysis.

Finally, the use of both PP and MCMC methods to explore the posterior distribution shows that the *shapes* of the respectively resulting *projected* and *marginal* distributions are near identical in log space, i.e., proportional, and therefore convey the same diagnostic information about the parameter space for most of the presented cases [21]. The main advantage of the projection methods, due to the deterministic exploration of the parameter space, is that the equipotential manifolds in the log posterior space caused by non-identifiable parameters do not affect the method's ability to obtain projections of the posterior [21]. The MCMC method's main advantages are computational efficiency, achieved by focusing exploration of the parameter space on regions of high posterior density, and also the possibility of utilising the sampled set of parameters $\theta_{[K]}$ to compute the posterior predictive distribution and marginal distributions for other parameters derived from the sampled θ [27,29]. Producing a stochastic forecast for the temperatures in the building could facilitate use of stochastic Model Predictive Control (MPC) [56,57], which also accounts for uncertainty in the calibrated model parameters.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

O.M. Brastein: Conceptualization, Methodology, Formal analysis, Writing - original draft. **A. Ghaderi:** Methodology, Writing - review & editing, Validation. **C.F. Pfeiffer:** Methodology, Writing - review & editing, Validation, Supervision. **N.-O. Skeie:** Methodology, Writing - review & editing, Supervision.

References

[1] D. Perera, C.F. Pfeiffer, N.-O. Skeie, Modelling the heat dynamics of a residential building unit: Application to Norwegian buildings, Model. Identif. Control 35 (1) (2014) 43–57, https://doi.org/10.4173/mic.2014.1.4.
[2] E.P.B.D. Recast, Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast), Off. J. Eur. Union 18 (06) (2010) 2010.
[3] M. Killian, M. Kozek, Ten questions concerning model predictive control for energy efficient buildings, Build. Environ. 105 (2016) 403–412.
[4] S.F. Fux, A. Ashouri, M.J. Benz, L. Guzzella, EKF based self-adaptive thermal model for a passive house, Energy Build. 68 (2014) 811–817.
[5] A.-H. Deconinck, S. Roels, Is stochastic grey-box modelling suited for physical properties estimation of building components from on-site measurements?, J. Building Phys. 40 (5) (2017) 444–471.

[6] P. Bacher, H. Madsen, Identifying suitable models for the heat dynamics of buildings, Energy Build. 43 (7) (2011) 1511–1522, https://doi.org/10.1016/j.enbuild.2011.02.005.
[7] T. Berthou, P. Stabat, R. Salvazet, D. Marchio, Development and validation of a gray box model to predict thermal behavior of occupied office buildings, Energy Build. 74 (2014) 91–100.
[8] H. Madsen, J. Holst, Estimation of continuous-time models for the heat dynamics of a building, Energy Build. 22 (1) (1995) 67–79.
[9] G. Reynders, J. Diriken, D. Saelens, Quality of grey-box models and identified parameters as function of the accuracy of input and observation signals, Energy Build. 82 (2014) 263–274.
[10] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics 25 (15) (2009) 1923–1929.
[11] L. Ljung, System Identification – Theory for the User, Prentice Hall information and system sciences series, Prentice Hall PTR, 1999.
[12] L. Ljung, Prediction error estimation methods, Circuit. Syst. Signal Process. 21 (1) (2002) 11–21, https://doi.org/10.1007/BF01211648.
[13] R. Johansson, System Modeling and Identification, Information and system sciences series, Prentice Hall, 1993.
[14] D. Di Ruscio, Combined Deterministic and Stochastic System Identification and Realization: DSR – A Subspace Approach Based on Observations, Model., Identif. Control 17 (3) (1996) 193–230, https://doi.org/10.4173/mic.1996.3.3.
[15] R. Ergon, D. Di Ruscio, Dynamic system calibration by system identification methods, in: European Control Conference (ECC), 1997, IEEE, 1997, pp. 1556–1561.
[16] N.R. Kristensen, H. Madsen, S.B. Jørgensen, Parameter estimation in stochastic grey-box models, Automatica 40 (2) (2004) 225–237.
[17] H. Madsen, Time series analysis, Chapman and Hall/CRC, 2007.
[18] R. Juhl, J. K. Møller, H. Madsen, ctsmr-Continuous Time Stochastic Modeling in R, arXiv preprint arXiv:1606.00242.
[19] R. Juhl, J.K. Møller, J.B. Jørgensen, H. Madsen, Modeling and prediction using stochastic differential equations, in: Prediction Methods for Blood Glucose Concentration, Springer, 2016, pp. 183–209.
[20] O.M. Brastein, R. Sharma, N.-O. Skeie, Sensor placement and parameter identifiability in grey-box models of building thermal behavior, in: Proceedings of The 60th Conference on Simulation and Modelling (SIMS 60), 13–16 August 2019, Västerås, Sweeden, Linköping University Electronic Press, 2009, p. tbd.
[21] A. Raue, C. Kreutz, F.J. Theis, J. Timmer, Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 371 (1984) (2013) 20110544.
[22] N.R. Kristensen, H. Madsen, Continuous time stochastic modelling, Mathematics Guide (2003) 1–32.
[23] T. Bohlin, S.F. Graebe, Issues in nonlinear stochastic grey box identification, Int. J. Adaptive Control Signal Process. 9 (6) (1995) 465–490.
[24] S.A. Murphy, A.W. Van der Vaart, On profile likelihood, J. Am. Stat. Assoc. 95 (450) (2000) 449–465.
[25] A.P. Dawid, Conditional independence in statistical theory, J. Roy. Stat. Soc.: Ser. B (Methodol.) 41 (1) (1979) 1–15.
[26] H. Pohjanpalo, System identifiability based on the power series expansion of the solution, Math. Biosci. 41 (1) (1978) 21–33, https://doi.org/10.1016/0025-5564(78)90063-9.
[27] J. Kruschke, Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan, Academic Press, 2014.
[28] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes in C++, vol. 3, Cambridge University Press, 2007.
[29] C. Bishop, Pattern Recognition and Machine Learning: All just the Facts 101 Material, Information science and statistics, Springer, 2013.
[30] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092.
[31] W.K. Hastings, Monte carlo sampling methods using markov chains and their applications, Biometrika 57 (1) (1970) 97–109.
[32] S. Rouchier, M. Rabouille, P. Oberlé, Calibration of simplified building energy models for parameter estimation and forecasting: stochastic versus deterministic modelling, Build. Environ. 134 (2018) 181–190.
[33] O.M. Brastein, B. Lie, C.F. Pfeiffer, N.-O. Skeie, Estimating uncertainty of model parameters obtained using numerical optimisation, Model., Identif. Control 40 (4) (2019) 213–243, https://doi.org/10.4173/mic.2019.4.3.
[34] O. Brastein, D. Perera, C. Pfeiffer, N.-O. Skeie, Parameter estimation for grey-box models of building thermal behaviour, Energy Build. 169 (2018) 58–68, https://doi.org/10.1016/j.enbuild.2018.03.057.
[35] H. Madsen, J. Holst, E. Lindström, Modelling non-linear and non-stationary time series, Lecture Notes, Technical University of Denmark, Dpt. of Informatics and Mathematical Modeling, Kgs. Lyngby, Denmark.
[36] M.A.S. Perera, B. Lie, C.F. Pfeiffer, Structural observability analysis of large scale systems using modelica and python, Model., Identif. Control 36 (1) (2015) 53–65, https://doi.org/10.4173/mic.2015.1.4.
[37] G.C. Goodwin, R.L. Payne, Dynamic system identification. Experiment design and data analysis, 1977.
[38] A.H. Jazwinski, Stochastic processes and filtering theory, Dover Publications Inc, 1970.
[39] I. Goodman, R. Mahler, H. Nguyen, Mathematics of Data Fusion, Theory and Decision Library B, Springer, Netherlands, 2013.
[40] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, Philos. Trans. R. Soc. London Series A, Math. Phys. Sci. 236 (767) (1937) 333–380.
[41] S. Kullback, A Note on Neyman's Theory of Statistical Estimation, Ann. Math. Stat. 10 (4) (1939) 388–390.
[42] A.E. Gelfand, S.K. Sahu, Identifiability, improper priors, and gibbs sampling for generalized linear models, J. Am. Stat. Assoc. 94 (445) (1999) 247–253.
[43] M.J. Bayarri, J.O. Berger, The interplay of bayesian and frequentist analysis, Stat. Sci. (2004) 58–80.
[44] O. M. Brastein, B. Lie, R. Sharma, N.-O. Skeie, Parameter estimation for externally simulated thermal network models, Energy Build. 191 (2019) 200–210. https://doi.org/10.1016/j.enbuild.2019.03.018.
[45] S. Rouchier, Solving inverse problems in building physics: an overview of guidelines for a careful and optimal use of data, Energy Build. 166 (2018) 178–195.
[46] A.N. Tikhonov, A. Goncharsky, V. Stepanov, A.G. Yagola, Numerical methods for the solution of ill-posed problems, vol. 328, Springer Science & Business Media, 2013.
[47] C. Van Loan, Computing integrals involving the matrix exponential, IEEE Trans. Automatic Control 23 (3) (1978) 395–404.
[48] D. Simon, Optimal state estimation: Kalman, H infinity, and nonlinear approaches, John Wiley & Sons, 2006.
[49] M. Bocquet, K.S. Gurumoorthy, A. Apte, A. Carrassi, C. Grudzien, C.K. Jones, Degenerate kalman filter error covariances and their convergence onto the unstable subspace, SIAM/ASA J. Uncertainty Quantif. 5 (1) (2017) 304–333.
[50] T.J. Rothenberg et al., Identification in parametric models, Econometrica 39 (3) (1971) 577–591.
[51] D. Venzon, S. Moolgavkar, A method for computing profile-likelihood-based confidence intervals, Appl. Stat. (1988) 87–94.
[52] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. 9 (1) (1938) 60–62.
[53] N. Cressie, Statistics for spatial data, vol. 4, Wiley Online Library, 1992.
[54] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, vol. 1, MIT Press Cambridge, 2016.
[55] M. Kuhn, K. Johnson, Applied predictive modeling, vol. 26, Springer, 2013.
[56] F. Oldewurtel, A. Parisio, C.N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, M. Morari, Use of model predictive control and weather forecasts for energy efficient building climate control, Energy Build. 45 (2012) 15–27.
[57] T.A.N. Heirung, J.A. Paulson, J. OLeary, A. Mesbah, Stochastic model predictive controlhow does it work?, Comput. Chem. Eng. 114 (2018) 158–170.