Bayesian Uncertainty Quantification and Emulation of Building Energy Models: a Tutorial

Dario Domingo[a,*], Mohammad Royapoor[b], Hailiang Du[a], Michael Goldstein[a]

[a]*Department of Mathematical Sciences, Durham University, Stockton Road, Durham, DH13LE, UK*
[b]*Mohammad, please fill the affiliation, , , ,*

Abstract

Correct calibration of building energy models is crucial to ensure that these can be reliably used to make predictions and inform design decisions. This work provides guidance and methodological steps for a systematic treatment of uncertainty before and during calibration. We introduce emulation, a Bayesian technique to build a fast statistical surrogate of the model. The emulator's speed allows to explore the model's parameter space extensively in short timescales. The emulator's predictions, alongside several quantified uncertainties, are then used to sequentially rule out model parameter configurations which are highly implausible to replicate the observed data. This procedure is known as History Matching (HM) and has been successfully employed even in high dimensions and with slow simulators. A package for its implementation is referenced. We discuss and illustrate HM on an example of a building energy model, where both energy consumption and temperature data are matched. The results show that only 0.3% of the model input space leads to outputs that can match the observed energy consumption. The percentage reduces to 0.01% if temperature constraints are also accounted for. The example shows the efficiency of HM to locate the small region where model outputs and observations can match, and therefore its ability to deliver models that lead to more robust decision support.

Keywords: Uncertainty, Model Discrepancy, Building Energy Models, History matching.

*Corresponding author
Email address: dario.domingo@durham.ac.uk (Dario Domingo)

1. Introduction

Up to 40% of primary energy use globally is attributed to buildings. Given the global decarbonisation efforts, accurate characterisation of building energy has become a key tool to enable optimal design, retrofit and investment decisions.

In order to produce high-quality modelling results that acknowledge modelling limitations, full treatment of uncertainty becomes essential. This has driven an accelerated trend in model uncertainty research and performance gap analysis. A recent article has reported that a large proportion of the building modelling community lacks essential knowledge on what the most fundamental parameter inputs for buildings are and how these impact model predictions [1]. This knowledge gap has major consequences, as retrofit strategies are mostly derived in consultation with building energy models. As such, techno-economic benefits of proposed solutions can be misleading. In addition to this, future buildings are expected to become more responsive to other civic activities (i.e. power generation and storage, transport, etc.)[2]. Therefore imminent horizons require aggregated energy data at district and potentially city level, where accuracy and treatment of uncertainty are invaluable tools for decision makers at all levels.

A preliminary step to support model-based decisions is model calibration. Calibration is the process through which values of the model parameters are chosen, so that model outputs replicate observations within prescribed tolerances. Classically, this is performed by running the model multiple times, for different values of the parameters. Hence, for each simulation, one or more measures of the discrepancy between the corresponding output and the observed data are evaluated. In the context of building energy consumption, ASHRAE guidelines [3] are often followed [4, 5] to identify error thresholds below which the discrepancy can be considered sufficiently small.

The main challenge of this approach is that the number of simulations needed to cover uniformly the model's parameter space grows exponentially with increasing parameter dimensions. To meet the computational challenges involved, a new area of Bayesian statistics has arisen, namely emulation [6–9], part of the wider field of uncertainty quantification (UQ) of computer models.

An emulator is a statistical surrogate of the simulator, providing predictions of the simulator's outputs at inputs for which the original model has never been evaluated. The emulator can be built and validated on a

personal device. It is instantaneous to run and therefore allows to predict the model results for a very large number of inputs. Moreover, it quantifies the uncertainty associated with such predictions. This explains the success of emulation in a wide range of fields where complex computer models are used, including but not limited to climate [10] and energy price projections [11], epidemiology [12], hydrocarbon reservoirs [13] and galaxy formation [14].

Emulation is developed under Bayesian principles, arguably the most natural framework to deal with uncertainties. In the context of calibration, key uncertainties that are often overlooked are: i) the ability of the model to simulate the process of interest (for example, building energy consumption); ii) and the intrinsic accuracy of the observations used to calibrate the model. To deal coherently with these and other sources of uncertainty, we propose the use of so-called implausibility measures to compare model (or emulated) outputs and observations.

Implausibility measures have been employed in the contexts of reconstruction of ice sheet shapes [15], gene evaluation [16], galaxy formation [14] and more, within a process known as History Matching (HM). The process sequentially removes regions of the input space deemed implausible to produce outputs which match the observations. However, in contrast to classical approaches in building energy model calibration, it does so in light of quantified uncertainties affecting various components of the system.

This work is aimed at researchers in the energy community, who are keen to make their model-based inference more robust, by operating in a framework that recognises and accounts for unavoidable sources of uncertainty. The statistical framework is illustrated on an example study of a single-house energy model, where model parameters are sought that simultaneously match energy and environmental features. A link to an R package is provided, which allows the researcher to implement the proposed methodology within their own research.

This work is organised as follows. Section 2 discusses key sources of uncertainty in computer model-based inference, particularly in the context of building energy models. Limitations of current approaches in this field are also outlined. Section 3 illustrates the proposed methodology: from emulation to implausibility measures and HM. Section 4 gives details of the building energy model and field data used in our case study. Results on the example study are illustrated in Section 5, where we also discuss approaches to scenarios that may be encountered in other case studies. Section 6 concludes the article with discussions and work extensions.

2. Uncertainty Sources in Building Energy Models

2.1. Calibration under Uncertainty

Calibration is a crucial topic not just in energy, but in a variety of scientific disciplines. It refers to the process by which the parameters of a computer model are identified so that model outputs match observed data. Mathematically, the model (or simulator) can be seen as a function f, with input $\boldsymbol{x}$ containing the list of model parameters of interest, and returning $f(\boldsymbol{x})$ as output. If we call z the observed data we want to match, the calibration task can be phrased as finding the input(s) $\boldsymbol{x}$ so that $f(\boldsymbol{x})$ is "close enough" to z. In practical applications, we often have a sequence of observations to match (and correspondingly a sequence of model outputs), rather than a single observation.

Proximity between simulations and observations should be assessed upon consideration of all sources of uncertainty that affect the system. While an exhaustive list of such sources is problem-dependent and challenging to specify in its entirety [7], two sources of uncertainty usually play a prominent role in energy system modelling and beyond:

1. *Model Discrepancy (MD).* Due to unavoidable modelling assumptions and numerical approximations, the output simulated by a model f at an input $\boldsymbol{x}$ will be different from the value that would be observed in the real world under the same physical conditions represented by $\boldsymbol{x}$. This difference is referred to as model discrepancy.
2. *Observational Error (OE).* It is a source of uncertainty intrinsic to all physical measurements, with a magnitude that depends on the employed measurement device

Accounting for these two uncertainty sources is crucial to assess whether a model has been successfully calibrated against observed data, and to make sure that robust inference can be drawn from the calibrated model predictions.

2.2. Model Discrepancy in Building Energy Systems

This section highlights sources of model discrepancy typical of building energy models, arising from a wide range of assumptions that analysts need to make

2.2.1. Building Thermal Properties

In a study of two office buildings in Australia [17], cooling set-points, lighting power density, ICT infrastructure and its schedule were found to impact the predictions of energy models the most. In other words, model predictions were found to be particularly sensitive to small changes in these parameters. Secondly, a major source of model discrepancy concerns the fabric thermophysical properties, the fabric composition (i.e., non-homogeneity, moisture content affecting thermal conductivity, etc.), as well as surface properties (radiative or convective characteristics). Most notably, the hypothesis of mono-dimensionality of heat flow - which is fundamental to thermal resistance (U-value) calculations in ISO 6946:2007 (Mohammed, can you please provide the bibtex references for SO 6946:2007 and ISO 9869-1:2014, corresponding to references number 5 and 8 of the word document?, ISO 9869-1:2014, CIBSE and ASHRAE [5-8] (same for CIBSE an ASHRAE, references number 6 and 7. I can't uniquely locate them on google scholar nor elsewhere.) (Maybe check, whether the ASHRAE reference - number 7 - can be replaced by my current number 3?) – remains a major simplification. In solid bodies, heat travels in a diffused and 3-dimensional manner which no building energy modelling platform replicates. Thirdly, for existing structures, notable differences in thermal property measurements exist across seemingly uniform building envelops [18–21]. These divergences are then compounded further given that they are used in U-value calculations (either semi-stationary or dynamic) that each produce a slightly different outcome [19] (disagreements of up to 393% are reported between measured and calculated figures [18]). While traditional solid masonry walls [22, 23] and floors [24] have been found to perform better than model calculations, actual composite walls are reported to perform worse than models suggest. A. Marshall et al. found CIBSE model U-Values for brick walls, ceilings and doors to over-predict performance by 30.3%, 15.5% and 9.9% respectively while calculated window properties showed much better matches with measured figures [25]. A study of 57 properties found that while variations between similar wall types and even within the same dwelling's walls existed, 44% of walls performed better than CIBSE calculations, 42% were within acceptable bounds and only 14% of sampled walls performed worse than calculations suggest [26], however measured and calculated floor U-values were found to be in good agreement. In summary difficulties in identifying the composite material and their density and moisture content, internal and external air velocities, fabric

non-homogeneity and cumbersome nature of in-situ fabric studies are notable reasons behind fabric U-value uncertainties.

2.2.2. Weather and Occupant Activity

The micro-climate around a building is by its nature random, with a particular set of conditions never repeating itself over the building's lifetime. It is affected by wind speed and direction, solar irradiance both outside and within the building, external temperature and humidity, cloud cover and ground albedo, atmospheric pressure. These all interact in close-to-chaotic manner to dictate a micro-climate, which is then represented in a simplified annual weather file for most energy simulations. Weather files are normally extracted from decades of actual data (e.g. using Finkelstein-Schafer method) but remain a major source of energy simulation uncertainty for the following reasons: i) building energy performance is affected by future rather than past weather; ii) a single weather file can hardly represent all meteorological conditions; iii) most weather files are not measured at the site but at a nearby meteorological station (often exposed airports), leading to inadequate representation of urban heat island and sheltering effects [27].

Second to micro-climate, the occupant behaviour is drastically simplified to inform the current generation of simulation tools. Temporal schedules usually describe the variations in occupant-related activities in a homogenous and deterministic form. More accurate representations of the occupant's dynamic interaction with the building remain an active area of research. Advances in pervasive sensing and data collection have enabled better understanding of occupant presence and movements. Insights from improved monitoring have informed two categories of occupant representation: implicit and explicit models that predict stochastic occupant interventions with windows, ICT and HVAC. Improvements are gained by seeking to understand the underlying logic behind occupant interaction with its surrounding at a behavioural level (explicit method) [28]. Challenges remain in adequately detailed monitoring of occupant behaviour, the integration of future occupant behaviour into simulation models and validation of the approach against longitudinal field studies. W. Tian et al. [29] note the need for further research to provide clear guidance not only to characterise stochasticity but also on the number of samples required to arrive at acceptable statistical power in building energy analysis [29]. This remains a priority particularly if post pandemic workplace attendance continues to evolve under a hybrid working regime.

2.3. Limitations of Current Calibration Approaches in Building Energy Models

None of the approximations and assumptions discussed in Section 2.2 undermines the validity of a model. However, awareness of similar assumptions and approximations is key to quantify the discrepancies induced when comparing model outputs and real-world observations.

In the context of building energy models, ASHRAE guidelines [3] are often followed to assess whether a model has been successfully calibrated against observed data [4, 5]. The guidelines make use of the following discrepancy measures between a sequence of N simulated model outputs S_i, and a corresponding sequence of observations M_i:

$$\mathrm{MBE} = \frac{\sum(M_i - S_i)}{\sum M_i} \tag{1}$$

$$\mathrm{CV(RMSE)} = \frac{\sqrt{\frac{1}{N}\sum(M_i - S_i)^2}}{\frac{1}{N}\sum M_i} \tag{2}$$

The model is considered calibrated if the relevant condition of the following two is met:

(a) Hourly measurements: $-10\% \leq \mathrm{MBE} \leq 10\%$ and $\mathrm{CV(RMSE)} \leq 30\%$.
(b) Monthly measurements: $-5\% \leq \mathrm{MBE} \leq 5\%$ and $\mathrm{CV(RMSE)} \leq 15\%$.

While the above criteria are easy to check, their use to assess model calibration presents some limitations. Firstly, acceptance thresholds are independent of the level of accuracy to which measurements are available, and of the level of discrepancy between the model and reality. Moreover, formulas (1) and (2) are not applicable to some of the physical quantities for which model outputs and observations may be available.

Indeed, the two formulas measure the magnitude of the (average) bias and root mean square error of the sequence $M_i - S_i$, relative to the (average) measured value. This is meaningful for any quantity intrinsically defined on a positive scale, such as energy consumption, but becomes less meaningful for other quantities, such as temperature. Here, using different units (*e.g.*, degrees Celsius or Kelvin) leads to seemingly different results. In reality, only Kelvin should be used, but in this case criteria (a) and (b) would be easily fulfilled due to the large denominator in (1) and (2).

In Section 3, we propose a statistical framework which overcomes these limitations. The framework makes it easy to account for recognised uncertainties when model and data are compared and it is applicable to any quantity of interest (and, in fact, to several quantities simultaneously). Moreover, thanks to the use of emulators, the framework can explore the model response at millions/billions of inputs, in a fraction of the time required for physics-based models to generate output.

3. Methodology

This section presents statistical methods to quantify and include different sources of uncertainty during model calibration. The overall process we illustrate is referred to as History Matching (HM). The name was first used within a study on hydrocarbon reservoirs [13] and indicates the attempt to "replicate history" i.e., to produce model outputs that match historical data. Since then, its statistical principles and methods have been successfully applied to a variety of fields (climate [15], biology [30], epidemiology [12], cosmology [14] and more). However, in energy systems there has been little to no application that are reported in the literature.

In this section we discuss the tools and statistical principles underlying HM, illustrate the procedures and provide statistical packages to implement it. We hope that our exposition, and the following illustration to a building energy example study, will promote its uptake within the energy community, and enable a robust treatment of uncertainty in calibration of computer models.

The HM procedure sequentially rules out regions of the model's input space where, in light of all quantified uncertainties, model outputs cannot match observed data. Emulators and implausibility measures are the two key tools used to accomplish the task as follows:

- Emulators allow to predict the model output at choices of input parameters where the model has never been run, quantifying the uncertainty of the prediction. Given their speed, emulators can be used to explore the model parameter space thoroughly, even in higher dimensions.

- Implausibility measures quantify the distance between model outputs and observations, in light of different sources of uncertainty.

Section 3.1 provides an overview of emulation, in simple but comprehensive terms. Sections 3.2–3.4 then formalise the concepts of model discrepancy

and measurement error, introduce implausibility measures, and illustrate the overall HM procedure, respectively.

3.1. Bayes Linear Emulators as Fast Model Surrogates

Evaluating the model's response across its input space, and comparing its outputs to observations, is a key part of calibration. However, running a number of simulations that thoroughly cover the parameter space proves unfeasible due to time and computational constraints. Emulators can be employed to overcome this issue.

An emulator is a statistical surrogate of the model, predicting the model's outputs at inputs where the model has never been run. Emulators' most evident advantages are speed and low computational requirements. A nominal measure of this speed is that predictions of the output of interest at thousands of different inputs can be obtained in one second on a personal laptop. Moreover, the emulator prediction is accompanied by a precise uncertainty statement of its accuracy, which lays the basis for model uncertainty analysis.

In this work we employ Bayes linear emulators (BLEs) to replicate the behaviour of the example-study building energy model (see section 4). One emulator is built for each output of interest. In this section, we give a brief overview of the choices behind the construction of a BLE, referring the reader to [6, 14] for further details. The interested reader is also referred to [31] for a much more in-depth treatment of Bayes Linear principles.

We use the following form for $f(\boldsymbol{x})$, the output of the computer model at input $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \sum_j \beta_j g_j(\boldsymbol{x}) + u(\boldsymbol{x})\,. \tag{3}$$

Equation (3) is made of two terms. The first one is a regression component: it uses a linear combination of known functions $g_j(\boldsymbol{x})$ to model the global behaviour of f across the input space. The second term, instead, is meant to capture the local fluctuations of f, once the main regression component has been accounted for.

Now suppose that the model has been run at a sequence of inputs $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$. Each $\boldsymbol{x_i}$ is called a design point and we will denote the corresponding output by $y_i = f(\boldsymbol{x_i})$. Then:

- The coefficient β_j in equation (3) can be estimated by fitting a linear regression model with predictors $g_j(\boldsymbol{x})$ to the known pairs $(\boldsymbol{x_i}, y_i)$.

- Values of the residual process $u(\boldsymbol{x})$ will then be known at each design point: $u(\boldsymbol{x_i}) = u_i$. These values will approximately oscillate around 0, with local patterns that the regression term has not been able to detect. We predict values of $u(\boldsymbol{x})$ via a BLE.

The idea to this last point is to model $u(\cdot)$ as a stochastic process, specifying a prior mean and covariance structure for it. In the Bayes linear framework, these are then adjusted to the observed values u_i at $\boldsymbol{x_i}$, yielding an adjusted mean prediction and uncertainty statement for $u(\boldsymbol{x})$, at any $\boldsymbol{x}$.

The prior specifications we make for u are as follows. We assume $u(\boldsymbol{x})$ to have mean zero at all inputs $\boldsymbol{x}$ and we use a stationary kernel to model the covariance between the value of u at any two inputs $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$. A stationary kernel is one for which the above covariance only depends on the difference $\boldsymbol{x} - \widetilde{\boldsymbol{x}}$. For common choices in emulation see Section 4.2.1 of [32]. In this work, we use the squared exponential kernel, leading to the following formula:

$$\mathrm{Cov}[u(\boldsymbol{x}), u(\widetilde{\boldsymbol{x}})] = \sigma_u^2 \exp\left(-\sum_k \left(\frac{x_k - \widetilde{x}_k}{d_k}\right)^2\right) \tag{4}$$

The subscript k in x_k and $\widetilde{x}_k$ denotes the k^{th} component of the two inputs, *i.e.*, the k^{th} model parameter. The positive coefficient d_k (called correlation length, or length scale) measures the strength of correlation in outputs when the k^{th} parameter is modified. Finally, the coefficient σ_u^2 denotes the prior variance of $u(\boldsymbol{x})$ at all inputs $\boldsymbol{x}$.

Expression (4) has been written in terms of all components of the input vector $\boldsymbol{x}$. However, just some of the parameters often prove relevant to explain the great majority of the variability in u. We call these active parameters and, in practice, only use them in expression (4). The small variability left in u, due to the inactive parameters, is modelled as uncorrelated noise with constant variance ν^2.

Notice that no distribution has been assumed for the process u. Within the Bayes linear framework, we only need to specify mean and covariance, as above. These mean and covariance are then adjusted to the values $u(\boldsymbol{x_i}) = u_i$, yielding an adjusted mean and variance for the value of $u(\boldsymbol{x})$ at any $\boldsymbol{x}$. Details of the adjusted mean and covariance formulas can be found for example in [6, 14, 31]. The above should however be sufficient to clarify the setting within which a BLE is built, and the decisions that need to be made. In particular, these are: the predictors g_j of the regression term; the variances σ_u^2 and ν^2; the active parameters; the correlation lengths d_k.

Although – put strictly - the above procedure builds a BLE for the "residual" process u, once the regression prediction at $\boldsymbol{x}$ is added, an overall prediction for $f(\boldsymbol{x})$ is obtained. Ultimately, this is the focus of interest. For this reason, in the following the term emulator will refer to the overall emulator of f. At any input $\boldsymbol{x}$, this provides a prediction $\hat{f}(\boldsymbol{x})$ of the unknown value $f(\boldsymbol{x})$, and a standard deviation $\hat{s}(\boldsymbol{x})$ quantifying the potential error of such prediction.

3.2. Model Discrepancy and Observational Errors

The concepts of model discrepancy (MD) and observational error (OE) are relevant to any uncertainty analysis linking computer simulations and measurements [13–16]. In this section we formalise them using a general framework. To relate the notation to the context of building energy models, the reader may think of: $\boldsymbol{x}$ as a set of model parameters; $f(\boldsymbol{x})$ as the simulated consumption in the building under parameters $\boldsymbol{x}$; y as the actual energy consumed in the building; z as the (imperfect) meter reading reporting the consumed energy.

Following[6], we assume that an appropriate choice of inputs exists, $\boldsymbol{x}^*$, that accurately represents the values of the system's parameters. We then link the simulator output $f(\boldsymbol{x}^*)$ at this parameters configuration and the real-world value y of the quantity being modelled, via the formula:

$$y = f(\boldsymbol{x}^*) + \varepsilon_{MD}\,, \tag{5}$$

where the model discrepancy term ε_{MD} accounts for the difference between the real and the simulated process. While the assumption is that just one quantity is being modelled, the same formulation can be considered if both y and $f(\boldsymbol{x})$ were vectors.

The unobservable real-system value y is usually estimated through a measurement z. We link these two quantities via the relationship:

$$y = z + e_{OE}\,, \tag{6}$$

where the term e_{OE} accounts for the observational error in the measurement.

The additive formulation in equations (5) and (6) is a simple but efficient way to model MD and OE, which also makes statistical inference tractable. Information about the quantities ε_{MD} and e_{OE} should in fact be sought in statistical form, rather than quantified as a single number.

Manufacturer guidelines are usually available to estimate the OE magnitude (e.g., up to 5% of the measured value). Estimates of MD magnitude may be more challenging to obtain. However, the modeller's knowledge of the assumptions and approximations used within the simulator, alongside literature research, usually provide guidance on the uncertainty effects of different assumptions and can therefore lead to an overall estimate of MD.

In the example study discussed in this work, we consider two different levels of MD (10% and 20% of the emulated value of $f(\boldsymbol{x})$), discussing how the choice affects the results (Section 5). For a more detailed treatment of MD, and the further distinction between internal and external MD, the reader is referred to [6].

3.3. Implausibility Measures

Given a computer model f of a physical system and measurements z of the system, this section outlines how to measure the likelihood that an input $\boldsymbol{x}$ to the model well represents the system's features. We do this by comparing $f(\boldsymbol{x})$ and z. As we will see, acknowledging the involved uncertainties will naturally lead to measure implausibility as opposed to likelihood.

The setting is as follows. A series of observations $z_1, \dots, z_m$ of m different quantities is available. Each quantity is modelled by the simulator: $f_j(\boldsymbol{x})$ denotes the simulator prediction of quantity j at input $\boldsymbol{x}$. For instance, in the case of a building energy model, z_j and f_j may represent measured and predicted consumption in month j. To make the notation lighter, in the following we initially drop the index j, reintroducing it later where information from different measurements/model outputs will be aggregated.

The construction of implausibility measures is based on the following reasoning. A choice of input parameters $\boldsymbol{x}$ well represents the hidden system's parameters (called $\boldsymbol{x}^*$ in Section 3.2) if the following two quantities coincide:

i) The value assumed by the real system under conditions $\boldsymbol{x}$ (eg, energy consumed under building parameters as in $\boldsymbol{x}$)
ii) The actual value y of the real system (eg, actual energy consumed in the building).

None of the two quantities is available for general $\boldsymbol{x}$, but both can be estimated, respectively by $f(\boldsymbol{x})$ (the model predicted consumption at $\boldsymbol{x}$) and z (the measurement of y, such as a meter reading).

Hence, by comparing the difference $f(\boldsymbol{x}) - z$ to the tolerance allowed by MD and OE, we can establish whether the quantities in i) and ii) may or

may not be the same. This leads us to consider the following quantity:

$$\frac{|f(\boldsymbol{x}) - z|}{\sqrt{\mathrm{Var}(\varepsilon_{MD}) + \mathrm{Var}(e_{OE})}} \tag{7}$$

High values of the the quantity above make the input $\boldsymbol{x}$ *implausible* to match the unknown system features $\boldsymbol{x}^*$ having led to observe z.

Notice that expression (7) requires knowledge of the simulated value $f(\boldsymbol{x})$. In practice, this is rarely readily available. However, if an emulator of f has been trained, the emulator prediction $\hat{f}(\boldsymbol{x})$ can be used as estimate of $f(\boldsymbol{x})$. To account for the substitution, the emulator variance $\hat{s}(\boldsymbol{x})^2$ should be included in the denominator of (7). Therefore, reintroducing the index j to indicate the particular output in question, we define:

$$I_j(\boldsymbol{x}) = \frac{|f_j(\boldsymbol{x}) - z_j|}{\sqrt{\mathrm{Var}(\varepsilon^j_{MD}) + \mathrm{Var}(e^j_{OE}) + \hat{s}_j(\boldsymbol{x})^2}} \tag{8}$$

the implausibility measure of input $\boldsymbol{x}$, with respect to quantity j. Finally, as overall measure of implausibility with respect to all the m quantities under consideration, we define:

$$I(\boldsymbol{x}) = \max_j I_j(\boldsymbol{x})\,. \tag{9}$$

The rationale behind the above definition is as follows. A high value of $I(\boldsymbol{x})$ implies that at least one of the $I_j(\boldsymbol{x})$ is high. In other words, the input $\boldsymbol{x}$ has led to an output $f_j(\boldsymbol{x})$ that is highly incompatible with the observed value z_j, even after accounting for the involved sources of uncertainty (MD, OE, emulation). In this case, we say that the input $\boldsymbol{x}$ is *implausible* to lead to a match. Vice versa, we call $\boldsymbol{x}$ *non-implausible*, since the involved uncertainties make it possible for $\boldsymbol{x}$ to lead to a match between model outputs and observations.

The history matching algorithm, described in Section 3.4, sequentially rules out all implausible inputs, identifying a final region where model outputs and observations can match in light of quantified uncertainties.

3.4. History Matching: The Algorithm

History matching (HM) proceeds in waves. The procedure is illustrated below and summarised thereafter in a schematic algorithm.

In the first wave, some or all of the model outputs for which observations are available are emulated. The implausibility measure (9) is computed across the space, and all inputs $\boldsymbol{x}$ for which $I(\boldsymbol{x})$ exceeds a given threshold are discarded as implausible. The remaining region, comprising the currently "non-implausible" points, is usually referred to as the Not-Ruled-Out-Yet (NROY) region.

The process is repeated, in consecutive waves. Crucially, at any new wave, additional simulations are run within the current NROY region, before new emulators are trained and an additional fraction of the region is discarded as implausible. The procedure terminates when the NROY region does not shrink any further, or it has become empty (all inputs have been ruled out as implausible).

The algorithm below summarises the steps. Comments on the procedure and guidance on some of the choices follow in Section 3.5.

History Matching Algorithm

Choose a positive threshold T (guidance provided in Section 3.5). Define a sequence of NROY regions $\mathcal{R}^0 \supseteq \mathcal{R}^1 \supseteq \ldots \mathcal{R}^k \supseteq \mathcal{R}^{k+1} \ldots$ as follows:

1. Initial Step ($k = 0$)
   (a) Let $\mathcal{R}^0$ be the whole input space. Perform a sequence of model runs in $\mathcal{R}^0$, and build emulators of some of the quantities for which observations are available.
   (b) Let I^0 be the implausibility measure in equation (9). Define the first NROY region as $\mathcal{R}^1 = \{\boldsymbol{x} \in \mathcal{R}^0 \, : \, I^0(\boldsymbol{x}) < T\}$.
2. Iterative Step ($k \geq 1$)
   (a) Run additional simulations within $\mathcal{R}^k$. Decide which outputs to emulate, and build emulators for them based on old and new runs within $\mathcal{R}^k$.
   (b) Compute the implausibility $I^k(\boldsymbol{x})$ over $\mathcal{R}^k$, and identify $\mathcal{R}^{k+1}$ as $\mathcal{R}^{k+1} = \{\boldsymbol{x} \in \mathcal{R}^k \, : \, I^k(\boldsymbol{x}) < T\}$.
3. Stop when $\mathcal{R}^{k+1} = \mathcal{R}^k$ or $\mathcal{R}^{k+1}$ is empty.

The flowchart in Fig. 1 illustrates the overall methodology.

3.5. Comments on the Procedure and its Strengths in Handling Uncertainties

We comment here on some of the choices needed to perform HM and on the overall strengths of the procedure, particularly in handling uncertainties.

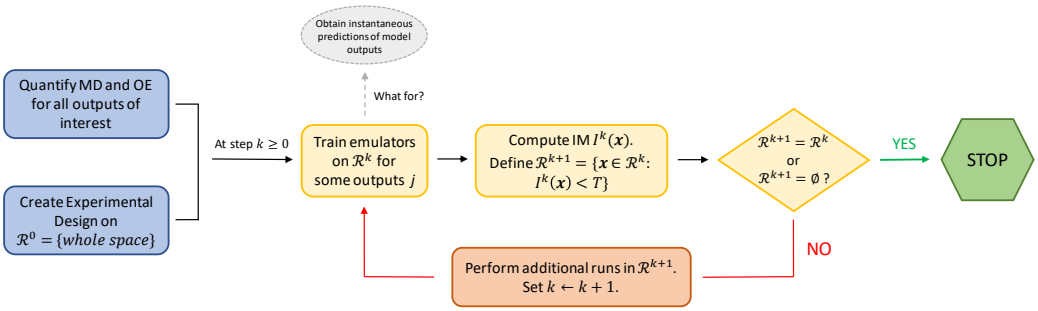

Figure 1: Flow chart illustrating the History Matching (HM) procedure. The initial steps (blue) consist in quantifying the main sources of uncertainty and designing a first set of runs over the whole space. The central part of the diagram (yellow) describes the steps of a typical wave, which aims to reduce the volume of the region comprising currently non-implausible inputs. Consecutive waves are repeated, till the region cannot be further reduced, or it is empty.

The choice of the threshold T reflects how large an implausibility we are prepared to accept, before ruling out an input as implausible. A common choice is $T = 3$, based on Pukelsheim's 3σ rule [33]. See [14] for more details. However, a larger threshold may be chosen if several quantities are being history-matched. Indeed, mainly due to emulation error, for each fixed output j there is some small probability that $I_j(\boldsymbol{x}) > T$ even if $\boldsymbol{x}$ was the "best" input $\boldsymbol{x}^*$. This probability increases if we consider the event that *at least* one j leads to $I_j(\boldsymbol{x}) > T$, that is, the event $I(\boldsymbol{x}) > T$. Therefore, a larger threshold T ensures that this probability of error is kept low.

As highlighted by point (2a) of the algorithm, at each wave, only some of the model outputs are used to compute the implausibility measure. This is a key feature of HM. Especially in early waves, some of the outputs may be difficult to emulate over large parts of the space, typically because the model's behaviour varies significantly across different input regions. As the search of non-implausible inputs is narrowed down to much smaller regions, outputs that were difficult to emulate may behave more uniformly within the region of interest. They can therefore be emulated precisely and included in the definition of $I(\boldsymbol{x})$ to rule out further implausible regions.

Also notice that, especially in earlier waves, not ruling out an input ($I(\boldsymbol{x}) \leq T$) does not suggest that the input leads to a match with the observations. Indeed, the implausibility of an input $\boldsymbol{x}$ may be low as a consequence

of large emulator uncertainty (term $\hat{s}_j$ in equation (8)), rather than actual proximity between the emulated prediction and the observation. However, emulator uncertainty is typically reduced between waves, because the emulators are trained on smaller regions (where the simulator may behave more smoothly) and on more design points. Reducing emulator uncertainty leads to larger implausibilities and, therefore, to more inputs being identified as implausible.

This is, in fact, the rationale behind proceeding in multiple waves and a key strength of HM. The decision of rejecting an output as implausible is taken in light of all quantified uncertainties at the given stage, including MD and OE. In this regard, notice that expression (8) can be easily adjusted to include additional sources of uncertainty in its denominator, whenever these are recognised and quantified in the problem in question. Moreover, by only involving differences and measurements of variability, implausibility measures can be computed for any quantity for which measurements are available, even if these are not measured on a positive scale (*e.g.*, Celsius degrees for temperature). This marks a crucial difference with respect to formulas (1) and (2).

In most cases, one or two waves are enough to rule out as implausible a great percentage of the space. In fact, as shown in the example study of Section 4, one wave may even be enough to identify the final non-implausible region, if the emulator uncertainties are negligible with respect to MD and OE. More generally, however, multiple waves may be needed. In R, the hmer package [34] can be used to implement emulation and HM. The package guides the researcher into multiple HM waves, by automatically suggesting a new design for each wave.

We conclude this section with a consideration. By designing model runs only where needed, HM allows the researcher to focus sequentially on the region of interest, while keeping the overall number of runs low. This is crucial in medium and high dimensions. In these cases, the region where a match is possible may only represent a tiny fraction of the original space. A more naïve yet not uncommon approach, where a number of simulations are run across the space and compatibility with observations is checked on these, would almost surely miss the region, even if the number of runs was very large. HM allows instead to identify the region precisely (even in high dimensions) thanks to the sequential "refocussing" procedure.

The event where, at some wave, all inputs have been discarded as implausible, is possible and indeed very informative. Causes for the lack of

mismatch between model and data may be multiple, with typical ones as follows: i) original parameter ranges are incorrect, ii) there is a problem with the data, iii) uncertainties are higher than estimated, iv) the model dynamics has a flaw. It is of course the researcher's task to step back and analyse what causes the mismatch, intervening on the model or in the assessment of uncertainties if necessary.

4. Example Study

This section describes the example study building used to illustrate the methodology outlined in Section 3. The building is a single, detached dwelling for which a model has been created in EnergyPlus (E+). Longitudinal energy consumption and temperature data is collected to calibrate the model.

4.1. Building and its Energy Usage

The detached two-story masonry construction used as case study was built in 1994. Two occupants are the only residents of the dwelling and were asked to archive their gas cooker and shower usage each day across an annual cycle. Given a very predictable pattern of occupancy (both occupants had 8am-5pm working commitments), it was possible to limit the stochastic nature of occupant activity as far as practically manageable and use deterministic schedules to represent occupant interventions with the building and its energy system. The building (with a gross area of 168.66m^2 and 19.73m^2 of unheated space) is located in a UK built-up urban surrounding and is only partly shaded on its west elevation by another adjacent property (shading represented in the model).

Across the monitoring year (2016) the property had an observed annual gas (15, 381 kWh) and electricity (2, 991 kWh) consumptions that respectively correspond to high and medium UK domestic consumption values [23] (Mohammad: reference 23, can you please provide it in bibtex format? Same with ref 24 and 25 below please. I cannot clearly identify them in google scholar). Occupants utilised shower facilities at a measured flow rate of 4.37 l/min and recorded on average eight 20-minute showers per week corresponding to an average of 50 l/person/day (occupants used cold water over wash basin and dishwasher supplied with cold feed only). These recorded values are below UK average domestic hot water usage (reported as 142 l/person/day [24] and 122 l/person/day [25]), but primarily reflect occupants' heavy use

of gym washing facilities. Gas cookers (containing 3kW and 5kW hubs) were used on average 4 times a week for 1 hour per cooking session.

4.2. The model: Co-dependency of Energy and Temperature Predictions

EnergyPlus (E+) is a collection of dynamic modules each simulating different environmental, climatic and operational conditions that define either the flow or the stored quantity of energy within building internal zones. The core of the programme is a heat-balance equation that is solved for all zones using one of three methods (3rd order backward difference, Euler method or analytically) to converge zone loads and resultant temperatures to within a pre-defined tolerance (using a predictor/corrector process). The uniqueness of E+ lies in it being a physically-based modelling solution. It oversees i) a simultaneous calculation of radiative and convective heat and mass transfer processes, ii) adsorption and desorption of moisture in building elements and iii) iterative interactions of plant, building fabric and zone air. This integrated and simultaneous simulation process is completed via several modules (and overseen by E+ simulation manager), with understandably multiple first-principle-based equations that are solved simultaneously and/or iteratively. This makes it very difficult to bring a sharp focus on any single or sets of expressions where model prediction uncertainties lie. However, the zone air heat balance equation is one of the primary mechanisms that describes the connected nature of heat gains/losses within a zone, the corresponding plant duty to offset them and the resultant zone mean air temperatures:

$$C_z \frac{dT_z}{dt} = \sum_{i=1}^{N_{sl}} \dot{Q}_l + \sum_{i=1}^{N_{sur}} h_i A_i (T_{si} - T_z) + \sum_{i=1}^{N_{zones}} \dot{m}_l C_p (T_{zi} - T_z) + \dot{m}_{inf} C_p (T_\infty - T_z) + \dot{Q}_{sys}\,, \qquad (10)$$

whereby $C_z\, dT_z/dt$ is the rate of change of thermal energy stored in the zone air, $\sum \dot{Q}_l$ is the sum of convective internal loads (Mohammad, can you check that in the first and third sums, the summing index is meant to be lowercase L?), $\sum h_i A_i (T_{si} - T_z)$ is convective heat transfer from the zone surfaces, $\sum \dot{m}_l C_p (T_{zi} - T_z)$ is the change of the room air enthalpy as a result of zone air mixing, $\dot{m}_{inf} C_p (T_\infty - T_z)$ is the infiltration heat transfer and finally $\dot{Q}_{sys}$ is the HVAC system input to achieve target temperature. Essentially each item on the right-hand side of equation (10) indicates a change of enthalpy

Figure 2: Left panel: power monitor used to characterise household appliances. Right panel: AC current sensor, monitoring transmitters and temperature sensors deployed in the case-study building.

due to environmental perturbations, while the left-hand side describes how these perturbations impact the zone air temperature and its enthalpy.

E+ assumes uniformity in i) zone air and surface temperatures, ii) incoming long/short-wave/diffused radiations and iii) surface properties (as opposed to treating radiation in a direct or point-based manner). It is reasonable to regard zone air temperature as the interconnection where conductive, radiative and convective heat balance and mass transfers are realised. This underpins our model validation approach in which energy and temperature data are essential.

4.3. Data Collection

A proprietary set of environmental and energy sensors were deployed to compile electricity and zone temperatures (Fig. 2). To reduce measurement uncertainty, each of the two target zones was equipped with two separate air temperature sensors at 1.3m above floor level and set to log data at 30s intervals to achieve a moderated average. Therefore, space temperature was recorded by 4 battery-powered sensors: two positioned in the south-facing master bedroom, and two in north-facing kitchen. Overall kitchen and master bedroom temperature sensors had total annual losses of 5.7% and 2.7% that required imputation. Each missing temperature cell was imputed by the average of the previous and successive available cells.

Table 1: Input parameter variations explored in the batch runs used to train our Bayesian emulators. Mohammad, many of these values need a careful checking between the two of us, see email :)

	Heating Setpoint (°C)	**Boiler Seasonal Efficiency (%)**	**Ext. wall U-value (Unit)**	**Roof U-val (Unit)**	**Floor U-val (Unit)**	**Infiltration rate (ach)**	**DHW Consumpt. (l/pers/day)**	**Cooking (%)**
Short name	V1	V2	V3	V4	V5	V6	V7	V8
Base-model Value	17.5	65	0.544 (out of range)	0.213	0.337 (out of range)	0.20	to be discussed	missing
Range	$[17.5, 20.5]$	$[60, 75]$	$[0.04, 0.063]$	$[0.15, 0.21]$	$[0.045, 0.055]$	$[0.2, 0.95]$	$[6.15 \times 10^{-6}, 2.2 \times 10^{-5}]$	$[1.05, 6.3]$

Gas consumption data was manually recorded at monthly basis using mains gas meter. Electricity consumption was logged at 10s intervals using two mains-powered clip-on current sensors on the incoming live cable providing two identical readings. The average of the two (in the end identical) readings formed the measured power usage. Gas and electricity data required no imputation.

In order to parameterise the energy model more accurately, a plugin power monitor was also used to characterise instantaneous and time-averaged consumption of the main electrical devices (TV, washing machine, ICT).

4.4. Model Input, Output and Range

By consulting manufacturers specification and the house builder's literature, a detailed set of parameter inputs for the model were compiled. Lower and upper bands were derived from scientific literature and used to dictate the size of associated variations explored in batch-runs (Table 1). Variations in internal and external air velocities are less pronounced in ground floors, but more significant for walls and roofs. Therefore, smaller floor uncertainty margin are reported in literature (Tables 1 and 2: Mohammad pls check correct referencing).

Manufacture's literature for glazing (installed in 2009) stated respective G and U-values of 0.69 and 1.79 W/m^{2}K. Respective error bands of $\pm$5%

Table 2: Parameter inputs for energy model development of the case-study building. (Dario-Mohammad: discuss caption and other red points in zoom call)

Parameter	**Description**
Heating	Natural gas boiler serving a radiator central heating system
Heating setpoint (setback)	19°C (16°C)
Heating schedule	02–11, 16–24
Ventilation	Natural ventilation (mechanical extract to family bathroom and en suite)
Ventilation rate	Highly stochastic, controlled by via openable windows
Gas boiler seasonal efficiency	65% (15 year old non-condensing gas-fired system boiler – 77°C/55∘C F+R)
DHW consumption	0.59 litre/m^2/day
Cooling setpoint (setback)	Uncontrolled
Nominal lighting power density	1.4 W/m^2 (manually controlled) to achieve 200 lux
Number of occupants	2 in total
Internal gains[1]	6 W/m^2
Gross (conditioned) area	168.66m^2 (148.93m^2)
Observed annua gas (electricity) consumption (2016)	15, 381 kWh (2, 991 kWh)
There were more, but are already present in Table 1	

and ±2% for G and U- values altered the gas consumption by ±2.05 kWh (±0.013%). Given its negligible nature, the error bands of the glazing were discounted in batch simulations. Given opaque fabric U-value impact on energy consumption [35], uncertainty bands are derived from literature and imposed on opaque U-values (Table 2).

Estimation of actual building infiltration rates requires convoluted air permeability tests. Table 4.16 of CIBSE guide A [6] (For ref 6 see one of above notes) outlines a range of 0.25 to 0.95 air change per hour (ACH) for various 2-story buildings below 500m^2 with a value of 0.5 ACH describing typical constructions similar to the case-study building. Therefore 0.5 ACH with a range of 0.25-0.95 ACH are used (meaning of 0.5? discuss with Mohammad). Local weather files compiled by a weather station approximately 3 miles away from the site was used to support the model development [27] (M: can you pls provide bibtex version of old reference 27).

We therefore consider eight uncertain parameters (Table 1), and use monthly gas and electricity consumption and hourly temperature data to history-match them. In the following, we will use the parameter to refer to any of the eight uncertain quantities in Table 1, and the term input to refer to a specific configuration of the eight parameters $\boldsymbol{x} = (x_1, \ldots, x_8)$.

5. Results

This section discusses the results of the application of the statistical principles and methods of Section 3, to the example study of Section 4. We build emulators of the model outputs of interest and history-match these outputs to identify a region of non-implausible inputs.

5.1. Experimental Design and Simulated Results

To build emulators of the model outputs, results of a small number of model runs are needed. An informal rule of thumb suggests using about 10 times as many runs as the number of varied parameters (eight in our case) [36]. In our case the simulator was moderately fast to evaluate, and a total of $n = 1,000$ simulations were run (taking about 16 hrs in E+). Note that this is substantially more than what is typically needed for eight inputs. While the results presented here are obtained by exploiting the information from all 1,000 runs, we have also simulated a more typical case where only 100 runs (randomly selected among the original 1,000) were used. This however yields no difference to the final results (see Section 6).

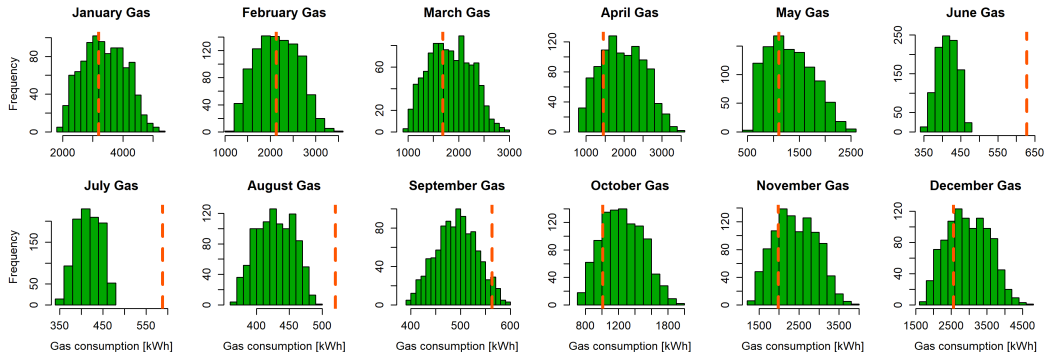

Figure 3: Distribution of simulated monthly gas consumption, for the experimental design used in this study ($n = 1000$ inputs). The dashed vertical line in a plot denotes the observed consumption for that month. Note that, for Jun-Jul-Aug, this lies outside of (and often far from) the range of simulated consumptions.

The experimental design (the set of n inputs for which the simulator is run) was chosen through Latin Hypercube Sampling (LHS). In two dimensions, the technique selects n points in a square so that exactly one point falls in each row and in each column of an $n \times n$ uniform grid of the square. The same idea governs higher dimensions. For this reason, LHS is commonly used in the design of computer experiment analyses [14, 37, 38], to identify points in a hypercube that fill the space well and are not too close to each other [39]. LHS is easily performed in R via the "lhs" package.

5.1.1. Gas

Fig. 3 shows simulated and observed values of monthly gas consumption for the $1,000$ design runs. A distinction between summer and non-summer months can be noticed: in June, July and August, all model runs considerably underestimate the observed consumption. This, in principle, does not rule out the possibility that other inputs within the explored space (an eight-dimensional hypercube with side ranges as in Table 1) may lead to a match. In this case, however, especially in June and July, it is evident that the range of simulated outputs is small with respect to the distance between these and the observed consumption: a simple linear model confirms indeed that a match within the explored space cannot be achieved for these months.

As illustration of the proposed methodology, in the following we will look for inputs that match all nine non-summer monthly gas consumption simultaneously, excluding the three summer conditions from the match. It is

important to note that in a real case study, the model's underestimation of summer consumptions should be identified and acted upon, prior to calibration/history matching. While the most likely reason for summer months deviation is the inability of deterministic schedules to capture stochastic aspects of energy governance in the household (*i.e.*, additional use of hot water/cooking in these months), our intervention with the model to reflect historical occupant behaviour for which we had little certainty would have been speculative. Greater insight into seasonal energy use variations would have allowed one or more parameters to be added as additional input variables to be history-matched.

5.1.2. Electricity

As described in Section 4.4, the model also predicts monthly electricity consumption. Within the explored ranges of the parameters, however, all months display almost no variability in output, with an observed consumption far outside the range of simulated values. For the three summer months (Jun–Aug), no variation at all is shown in simulated consumption among the n runs. This is partly an outcome of the deterministic electrical load specification. More importantly though, this is due to the fact that any changes with the 6 variables in Table 1 except for heating and DHW (V1 and V7) will only affect gas consumption. This makes the variability of electricity consumption very limited and bound to a small deterministic input schedule (e.g., a fixed 6 W/m2^2 for internal gains as dictated by field measurements).

Given the illustrative role of the example study within this work, we will not attempt to match the electricity consumptions in the methodological illustration that follows. We will instead consider the nine conditions coming from the observed monthly gas consumption in non-summer months, to which we will add temperature constraints in Section 5.5.

5.2. Emulation of Monthly Gas Consumption

For each of the nine non-summer months, the $n = 1,000$ simulations provide a dataset of n pairs $(\boldsymbol{x_i}, y_i)$ where $\boldsymbol{x_i}$ represents one configuration of the 8 input parameters, and y_i is the simulated gas consumption associated with it. We use the dataset to build an emulator of that month's gas consumption. To that aim, the eight input parameters are linearly rescaled so that each of them ranges in the interval $[-1, 1]$.

From Section 3.1, recall that several choices (on covariates, correlation lengths, prior variances) have to be made when building an emulator. To

validate these choices, we split the dataset as follows:

- Training set (700 runs): used to train the emulators.
- Evaluation set (150 runs): used to decide on the values of the emulator hyperparameters, by comparing the emulator's performance on this set to the known simulator's outputs.
- Test set (150 runs): used to test the previously built emulators on a completely new set of runs not used for training and evaluation.

Following the notation of Section 3.1, the prediction of the simulated consumption at an input $\boldsymbol{x}$ is computed as sum of: i) a linear regression part, ii) a prediction of the regression residual at the point $\boldsymbol{x}$, in the form of a Bayes linear emulator. Details on each of the two parts follow.

5.2.1. Linear Regression

The only choice to be made in building a linear regression model concerns the predictors to use, *i.e.*, the functions g_i in equation (3). For all the months of interest, a preliminary exploration reveals that the response y (gas consumption) is very well explained as quadratic function of the eight input parameters, which in the following we call $V_1, \ldots, V_8$ for convenience. Thus, we proceed as follows.

Let $\mathcal{P}$ be the set of all mutually orthogonal linear, quadratic and interaction terms of $V_1, \ldots, V_8$.[2] For a given integer k, consider the linear model with highest adjusted coefficient of determination (adj. R^2) and consider the smallest integer k for which the above linear model with k predictors has an adj. R^2 exceeding the threshold.

In our case, 10 predictors yield models with notably high adj. R^2, about 0.999 across all months. To keep the approach uniform among months, we consider the linear model with the "best" 10 predictors for each month (Table 3). As pointed out above, however, a different number of predictors for each output may be considered on other case studies.

Note that, in general, such a high R^2 should raise concerns of overfitting. In our case, the concern can be ruled out by observing that only 10

[2]Obtained in `R` via the command `poly(X, deg=2)`, where `X` is the 700×8 matrix whose rows are the training inputs. For 8 parameters, there is a total 44 linear, quadratic and interaction terms.

Table 3: Properties of the gas consumption emulators. For each month, from left to right: covariates used to build the linear regression model; adjusted R^2 of the linear model; variance of the residuals; active parameters used in the covariance function; value of the correlation lengths (same for all active parameters). In the predictor's column, the * symbol denotes the combination of all linear and interaction terms: $a * b * c = \{a, b, c, ab, ac, bc\}$.

	Predictors	**Adj. R^2**	**σ^2**	**Act. Params**	**d**
Jan	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9998	85.94	V_1, V_2, V_3, V_6, V_8	0.8
Feb	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9997	52.30	V_1, V_2, V_3, V_7, V_8	0.65
Mar	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9997	50.14	$V_1, V_2, V_3, V_6, V_7, V_8$	1.3
Apr	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9998	50.04	V_1, V_2, V_3, V_6, V_8	1
May	$V_1 * V_2 * V_6, V_3, V_8, V_1^2, V_6^2$	0.9989	206.45	$V_1, V_2, V_3, V_4, V_6, V_8$	1
Sep	$V_1 * V_2 * V_6, V_3, V_4, V_8, V_6^2$	0.9994	0.91	$V_1, V_2, V_3, V_6, V_7, V_8$	1.2
Oct	$V_1 * V_2 * V_6, V_3, V_4, V_8, V_6^2$	0.9996	24.40	$V_1, V_2, V_3, V_6, V_7, V_8$	1.4
Nov	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9998	57.35	$V_1, V_2, V_3, V_6, V_7, V_8$	1.2
Dec	$V_1 * V_2 * V_6, V_3, V_4, V_2^2, V_6^2$	0.9998	68.77	$V_1, V_2, V_3, V_6, V_7, V_8$	1.3

predictors have been used to explain 700 observations. The high coefficient of determination mirrors an intrinsically quadratic model dynamics.

5.2.2. Emulators of the Residuals

To build an emulator of the regression residuals of each month's consumption, choices about the following quantities are to be made: active parameters $\boldsymbol{x_A}$, correlation lengths d_i, prior emulator variance σ_u^2 an nugget variance ν^2 - see equation (4) and text thereafter. We proceed as follows.

- **Active parameters $\boldsymbol{x_A}$**: To identify them, we look at the most significant (in the sense discussed in Section 5.2.1) second- and third- order terms in a linear regression model of the residuals. Variables appearing by themselves with a high t-value ($t > 8$) are included as active parameters. The inclusion of variables appearing alone with a lower t-value or in interaction with other variables is instead considered case by case, according to the emulator performance on the evaluation set – see Section 5.2.3.

- **Correlation lengths** d_i: Once the active parameters are chosen, the same correlation length d is used for all of them. The value of d at each month is chosen by assessing the emulator's performance on the evaluation set and is reported in Table 3. Notice that, to attain a similar level of correlation across the space, higher correlation lengths are used when a higher number of active parameters is present.

- **Prior variances σ_u^2 and ν^2**: Let σ^2 denote the variance of the regression residuals being fitted. Hence, we set $\sigma_u^2 = 0.95\sigma^2$ and $\nu^2 = 0.05\sigma^2$.

Table 3 provides details of all choices made to build each of the nine emulators, including the ones concerning the regression line. We discuss validation of the emulators in the next section.

5.2.3. Emulator Validation and Performance

The active parameters $\boldsymbol{x_A}$ and correlation lengths d in Table 3 are chosen based on the emulator's performance on the evaluation set. The latter consists of 150 pairs $(\boldsymbol{x_i}, y_i)$ not used in the emulator's training, where each y_i is the simulated output at input $\boldsymbol{x_i}$. At each such $\boldsymbol{x_i}$, the emulator provides a prediction $\hat{y}_i$ of the simulated output and an uncertainty statement about

the prediction in the form of a standard deviation $\hat{\sigma}_i$. One way to assess the emulator's performance at $\boldsymbol{x_i}$ is thus to define:

$$\varepsilon_i = \frac{\hat{y}_i - y_i}{\hat{\sigma}_i} \,. \tag{11}$$

We call ε_i the emulator standardised error at $\boldsymbol{x_i}$. It represents the number of standard deviations which separate the emulator prediction from the known simulated output.

As we make no distributional assumption on the emulator, we can appeal to Pukelsheim's 3σ rule [33] to constrain expected values of ε_i: the result states that, for any continuous unimodal distribution, at least 95% of the probability mass lies within 3 standard deviations from the mean. Therefore, by only assuming unimodality of the emulator distribution, we should expect at least 95% of the ε_i to lie between -3 and 3.

For each of the nine emulators, we consider the plot of ε_i versus the predictions $\hat{y}_i$. Given a good emulator, such a plot should be characterised by an approximately random scattering of the points around the line $\varepsilon = 0$, with about 95% of them in modulus less than 3 due to Pukelsheim's rule. We assess visually both these properties for different choices of active parameters and correlation lengths, choosing the ones which return plots with the desired properties.

We tend to be slightly conservative in this phase, by choosing correlation lengths which generally yield more than 95% of $|\varepsilon_i|$ less than 3 . This is to prevent making choices tailored to the specific points used for validation. Once the parameters $\boldsymbol{x_A}$ and d are chosen, we compute the standardised errors in equation (11) on the 150 elements of the test set, to check that no anomaly shows up on a set of points never used in training or evaluation.

Finally, we note that the emulators of the nine monthly gas consumptions are remarkably precise in their predictions. This is a consequence of both an excellent regression fit to the data, and of a further reduction in the residual uncertainty due to the emulators. Quantification of this reduction is provided in Table 4, which shows the original variance of the residuals and an empirical 95% confidence interval (CI) of the emulator variance on the 150 test points. The CIs show that the uncertainty in the emulator predictions is very small compared to the range of simulated outputs (Fig. 4).

Table 4: For each month: variance of the regression residuals to which the emulator is fitted (first row); empirical 95% confidence interval of the emulator variance on the 150 test points (second and third row).

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Sep**	**Oct**	**Nov**	**Dec**
$\boldsymbol{\sigma}^2$	85.9	52.3	50.1	50.0	206.4	0.91	24.4	57.4	68.8
[**2.5%,**	6.6	5.9	3.1	3.2	16.8	0.06	1.5	3.8	4.3
[**97.5%**]	31.0	28.0	9.3	10.0	68.9	0.20	3.8	12.8	12.7

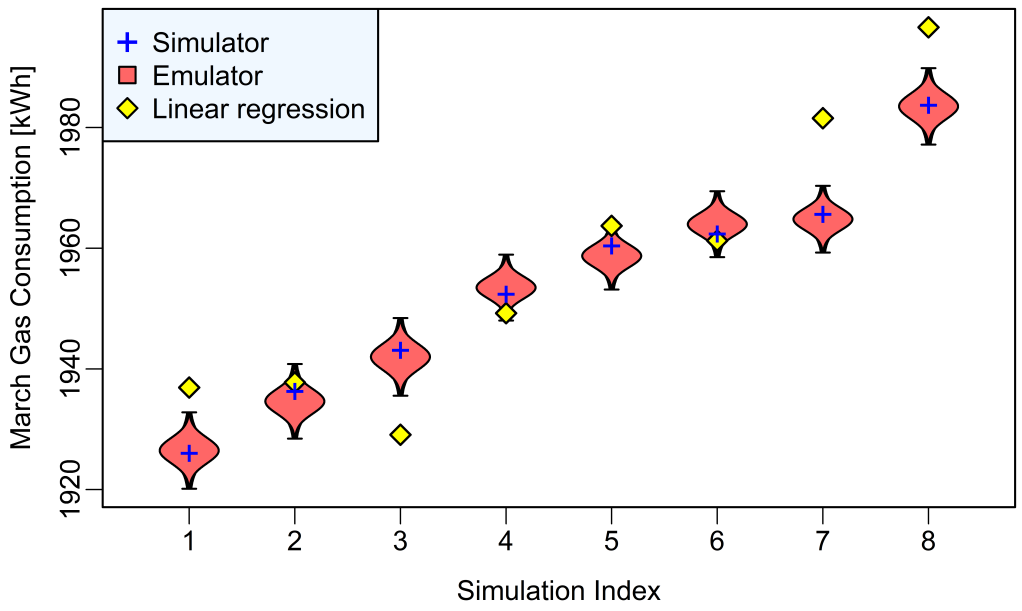

Figure 4: Comparison between point regression predictions and emulator predictions with associated uncertainty, for a sequence of randomly selected eight consecutive simulated March gas consumptions of the test set. The bell-shaped curves cover a range of 3 emulator standard deviations up and down the emulator prediction; the simulated value (+) always lies within this range, while the regression prediction may lie substantially further. Results are displayed for eight simulations to have a small range in the y-axis and allow to appreciate detail.

5.3. Observational Error and Model Discrepancy

The validated emulators can be used as probabilistic surrogate of the original model, to predict the monthly gas consumption at any input $\boldsymbol{x}$. We use the implausibility measure described in Section 3.3 to compare the prediction with the observed consumption z. The observational error (e_{OE}) and model discrepancy (ε_{MD}) are set as follows:

1. The OE is set to 5% of the observed value z, in agreement with the largest manufacturer's accuracy band ($\pm$5% for power monitor).
2. For illustrative purposes, we set MD to either 10% or 20% of the emulated consumption and compare results in the two cases.

Note that, as discussed in Section 2.2, accurate estimation of model discrepancy requires careful statistical analysis and possibly additional model runs. However, an order of magnitude between 10 and 20% of the simulated values is likely to represent a good estimate in many energy applications. If there are reasons to believe that model discrepancy is much higher, then the possibility of revisiting the model itself should be considered, by possibly including in the model key factors affecting the simulated dynamics.

5.4. History Matching

The HM procedure has been described in Section 3.4. In our case, at an input configuration $\boldsymbol{x}$:

1. We first compute the implausibility measures $I_M(\boldsymbol{x})$ for months M, as in equation (8);
2. We then define the overall implausibility $I(\boldsymbol{x})$ as in equation (9), $I(\boldsymbol{x}) = \max_M I_M(\boldsymbol{x})$;
3. Hence, we consider an input as non-implausible if $I(\boldsymbol{x}) < 4$.

As explained in Section 3.5, choosing a higher threshold than the more common $T = 3$ ensures that the probability of incorrectly rejecting an input $\boldsymbol{x}$ as implausible is kept small, whenever several observations are being matched simultaneously (nine in our case). Also notice that an alternative approach may be to define $I(\boldsymbol{x})$ as the second- or third-highest value of $I_M(\boldsymbol{x})$ over all months, while keeping a relatively low threshold T. However, if a particular month proves more challenging to match, this approach risks to classify several points as non-implausible, although those points are highly unlikely to match the observed consumption for that month.

In our example study, the percentage of space classified as non-implausible after one wave is 0.30% when 10% MD is used, and 19.52% for 20% MD. These percentages have been estimated by computing the proportion of non-implausible points on a sample of size $N = 10^7$, generated via a Sobol sequence in the eight-dimensional unit hypercube. The relative error on these estimates (approximately equal to $[(p(1-p)/N]/p$, where p is the estimated fraction) is therefore order of 10^{-2} and 10^{-3} in the two cases respectively. This makes both estimates very accurate.

In this case, we do not need to proceed to further waves. For all the months, the emulator uncertainty (Table 4) is already 1-2 orders of magnitude lower than the combined one from MD and OE. Additional waves therefore leave the implausibility measure essentially unchanged. We comment further on this in Section 6.

5.4.1. Visualisation of the Non-Implausible Region

The non-implausible region lives in an eight-dimensional space, one dimension per input parameter. To identify variables which play a role in constraining the region, we look at two-dimensional scatter plots of the non-implausible points, for all possible pairs of the 8 variables. For brevity, we do not show these all here. However, in the 10% MD case, the plots reveal that two variables are particularly significant: infiltration rate (V_6) and cooking energy consumption (V_8). One additional variable (heating setpoint, V_1) also seem to play a role in identifying non-implausible points.

Fig. 5 contains minimum-implausibility (MI) and optical-depth (OD) plots for all pairs of the three above variables. The MI plot of a pair of parameters shows the minimum value of the implausibility measure $I(\boldsymbol{x})$ over the hidden six dimensions. Thus, values greater than $T = 4$ in a MI plot identify pairs of the two parameters in question that will always be classified as implausible, irrespectively of the value taken by the remaining six parameters. For values smaller than $T = 4$, the OD plots instead show the fraction of non-implausible points, in the hypervolume extending over the six hidden dimensions.

Fig. 5 reveals that only values of the gas cooking higher than 5% of total household energy demand are able to yield a match with the observed monthly consumptions. Moreover, within the explored ranges, such values should be paired with values of the infiltration rate (right panels) approximately between 0.25 and 0.65 ach uniformity in lower/capital ach. Mohammad, preference?). Note that the two variables seem to be relatively

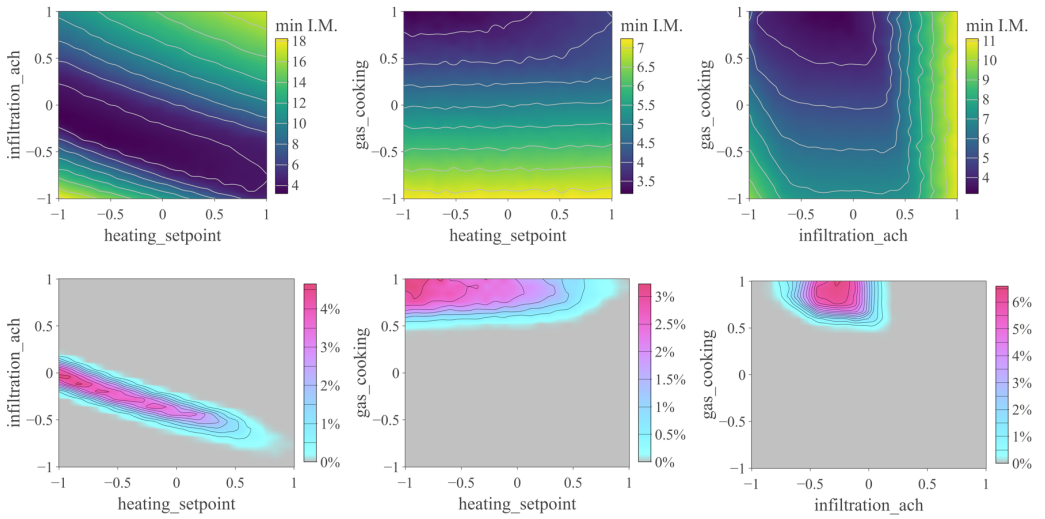

Figure 5: Plots of the non-implausible space when 10% MD is considered. Each panel in the top row shows the minimum implausibility along the six dimensions not included in the plot. Panels in the bottom row show the percentage of non-implausible region along the hidden dimensions.

independent. A stronger dependence can be seen instead for non-implausible values of heating setpoint and infiltration rate (left panels). Heating setpoint values are by themselves only minimally contrained: however, higher infiltration rates yield more constrained, lower values of the heating setpoint.

The left subplots in Fig. 5 also suggest that non-implausible values of the heating setpoint may as well be found to the left of the range originally deemed appropriate for this input variable. A similar consideration is valid for gas cooking values higher than the relevant range. Similar findings are not uncommon, especially during a first wave of HM. In such a case, it is advisable to step back and expand the ranges in question, running additonal simulations in the new region before proceeding with HM. However, for brevity purposes and in accordance with the methodological aim of the work, in this illustrative example we limit our discussion to the hypercube with ranges specified in Table 1.

5.4.2. Sensitivity to Model Discrepancy Magnitude

The plots in Fig. 5 concern the case where the magnitude of model discrepancy is 10%. Fig. 6 shows MI and OD plots for the same three variables, in the case of 20% MD. Roughly, similar patterns to the ones of Fig. 5 emerge,

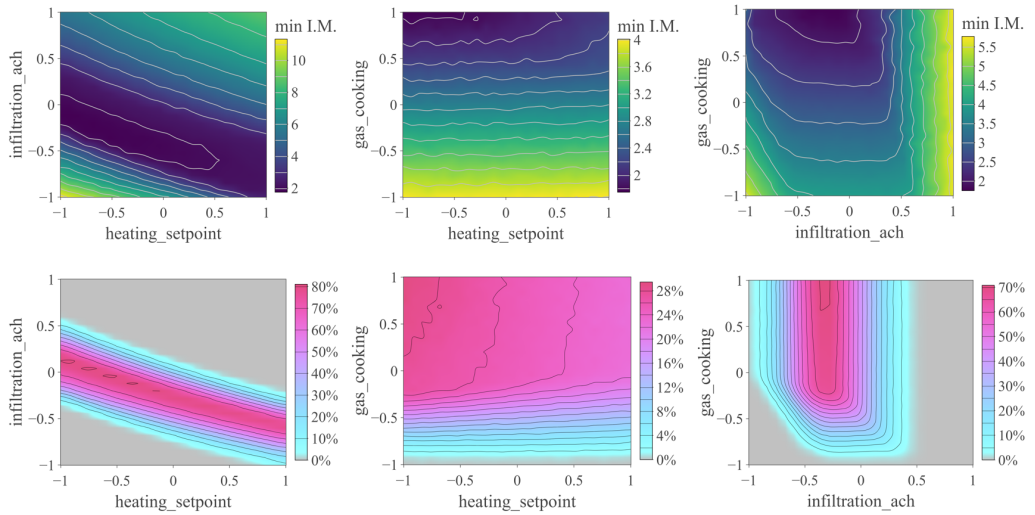

Figure 6: Same content as in Fig.4, when 20% MD is considered. Each panel in the top row shows the minimum implausibility along the six dimensions not included in the plot. Panels in the bottom row show the percentage of non-implausible region along the hidden dimensions.

albeit spread over larger regions: due to a lower confidence in the model, we rule out fewer points as implausible. The percentage of non-implausible points has risen to 19.52%, from only 0.3% in the 10% MD case.

This comparison highlights the potential sensitivity of history matching results to the choice of MD. Note that the precision of the emulator(s) used also plays a role in this. In our case, as Table 4 shows, we have remarkably precise emulators. Hence, essentially all the uncertainty accounted for in comparing emulator predictions and observations comes from model discrepancy and, to a lower extent, measurement error. Doubling the MD will thus make points significantly more likely to be deemed non-implausible. This once more highlights the importance of assessing the right order of magnitude of model discrepancy prior to performing history matching, as discussed in Sections 2.3 and 3.2.

5.4.3. Role of Different Constraints and Correlation among Outputs

The emulators allow to explore a wide range of questions, for which it would otherwise be impossible to draw sound inference. We briefly discuss one such example here, by looking at the compatibility between the imposed constraints (observed consumptions) and the correlation of model predictions

across different months. Note, in fact, that a strong correlation between two model outputs restricts the pairs of observed consumptions that can lead to a match.

Each of the upper-diagonal panels of Fig. 7 shows a scatter plot of emulated gas consumption of the relevant pair of months, on a random sample of $5,000$ inputs. The observed consumption to be matched is identified by a red cross. In most cases the latter is outside the region of outputs. However, the presence of model discrepancy (10%) and observational error allows to identify a region of non-implausible outputs, which is highlighted in turquoise in each plot. Note, this is a subregion of the output space, not of the input space as in Figures 5 and 6. A zoom of this region is displayed in the lower-diagonal panels, with points coloured by the overall IM (9). The observations aimed to be matched are displayed within a shaded rectangle identifying 5% observational error.

Finally, the histograms on the diagonal of the same figure show the distribution of the IM of a given month, on the points which would be classified as non-implausible if that month was not considered. The IM is reported with its orginal sign (expression (8) without absolute value), with positive values denoting an emulated consumption higher than the observed one. Values outside the range $[-4, 4]$ are highlighted by a darker colour. Months such as January, March, November and December, where all values are between -4 and 4, do not contribute to reduce the space once the other eight months' constraints are included. On the other hand, a month such as September rules out around 85% of the space that would be considered non-implausible without it.

As the upper-diagonal panels reveal, some of the simulated monthly consumptions are highly correlated, *e.g.*, January's and February's. In similar cases, imposing that one month's observed consumption is matched will limit which observed consumption of the other month can also be matched. In the case of January and February, the location of the cross is very close to the line of simulated consumptions, hence both observations can be easily matched simultaneously (and can also be matched at the same time of all other seven months, as we have seen). However, despite their high correlation, the two months do not play an interchangeable role. The January constraint plays a redundant role the February one is accounted for. But the opposite is not true, as the x-range of the two diagonal histograms reveal. The same is true for other pairs of strongly correlated outputs (*e.g.*, Feb-Mar, Feb-Nov, Oct-Nov to a lower extent).

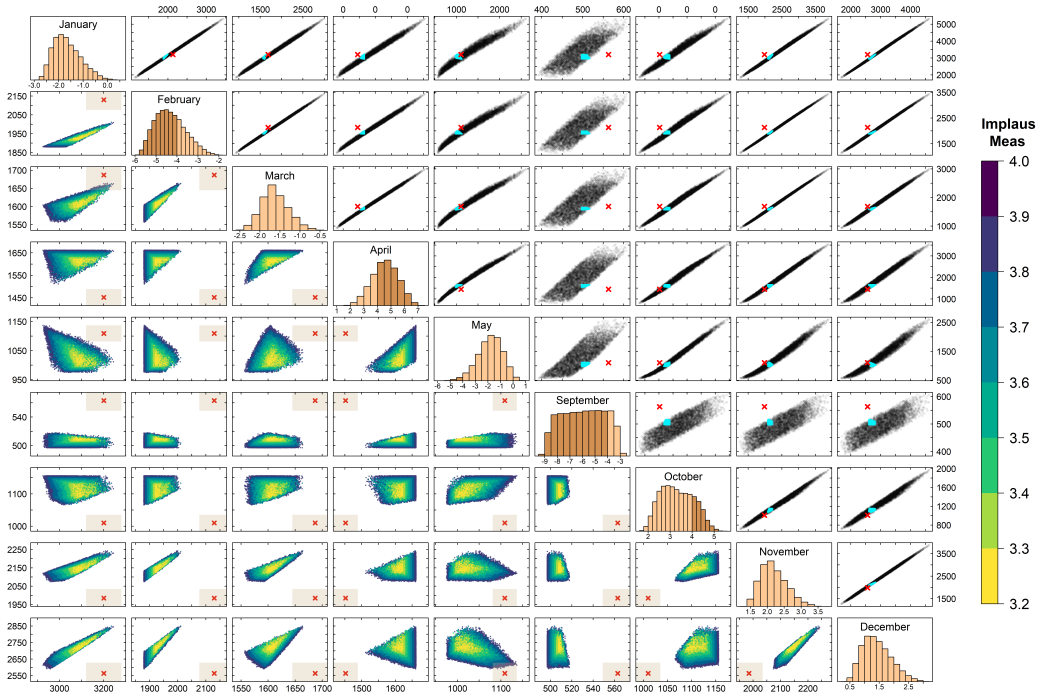

Figure 7: Information on model outputs, observations, and each month's contribution to cutting the non-implausible region. Upper-diagonal panels: scatter plot of emulated gas consumption for each pair of months (randomly selected 10,000 inputs). The red cross locates the observation to be matched, the light blue stain locates the non-implausible region. Lower-diagonal panels: zoom of the non-implausible region, coloured by implausibility measure (IM). Shaded rectangle around the cross identifies 5% measurement error. Diagonal panels: distribution of each month's IM, on the space deemed non-implausible when only constraints from the remaining months are considered. The darker colour denotes values outside the interval $[-4, 4]$, *i.e.* identifies inputs which transition from being non-implausible to being implausible when that month's constrain is added.

Finally, we notice that many of the lower diagonal projections appear delimited by straight lines. These generally represent the "boundary" beyond which one of the month's IM exceeds 4 in absolute value. This is particularly evident for the months whose diagonal histograms extend beyond the interval $[-4, 4]$. In this case, the non-implausible region is "cut" perpendicularly to that month's axis, on the lower or higher side of the range according to whether the histogram exceeds -4 or 4.

5.5. Adding Temperature Constraints

Implausibility measures were previously used to history-match observed gas consumption. They can however be used with any model output for which observations are available, a feature that marks a difference with measures such as MBE and CV(RMSE), as discussed in Section 2.3. In our case, we can history match the model temperature predictions in the kitchen and master bedroom, for which hourly time series are available both as observations (field data) and as model outputs.

Accounting for uncertainty when comparing simulations and observations is more challenging for time series than for scalar quantities. Appropriate tools may involve accounting for correlation across time, dimension-reduction techniques (*e.g.*, PCA) and multidimensional implausibility measures. It is beyond the scope of this work to go into these details, some of which are active areas of statistical research. However, scalar quantities may be extracted from a time series and history-matched through the same methodology discussed in Section 3. We show an example here.

In order to include summer constraints, we consider the following scalar quantity: the average temperature difference in July between day (8am–11pm) and night (00am–7am) hours. For each of the two rooms, we compute the above quantity across the $1,000$ design simulations, build an emulator of it as a function of the 8 input parameters, and history-match the parameters.

Notice that the choice of matching temperature difference introduces a different type of constraint to the one imposed on gas, where absolute consumption was matched. We consider the same levels of MD and OE as in Section 5.3. Accounting for each of the two temperature constraints separately rules out as implausible 17.96% (kitchen) and 24.30% (master) of the space respectively (we use here the threshold $T = 3$ since only one condition was considered in each of the two cases). Unsurprisingly, we note a strong dependence between the constraints coming from the two rooms: 12.38% of the space is classified non-implausible with respect to both constraints at

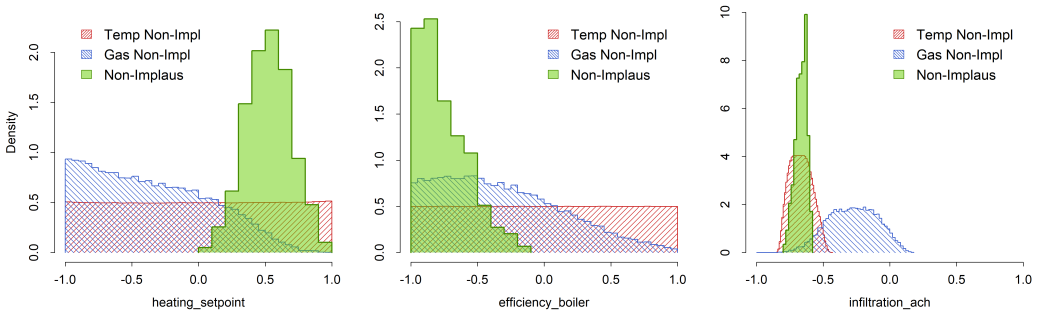

Figure 8: Marginal distribution of three input variables on non-implausible points. The shaded histograms mirror non-implausibility with respect to gas or temperature constraints only, the filled histogram with respect to both.

the same time, which is almost three times more than expected if they were independent.

Accounting for both energy (gas) and environmental (temperature) constraints at the same time classifies only 0.00975% of the original eight-dimensional cube as non-implausible. This is about four times less than it would be expected if the two constraints were independent. The interplay between energy and environmental constraints is highlighted in Fig. 8, in the case of three input variables. For each one of them, the plot shows the distribution of non-implausible values for that variable, when the constraints from either gas, or temperature, or both are considered.

As illustration, we comment here on the left plot of Fig. 8, which concerns non-implausible values of the heating setpoint (HS) parameter.

On the one hand, if we only account for the July temperature constraint, the original uniform distribution of heating setpoint values remains unchanged (red shade): this is expected, as heating is switched off in summer and does not, therefore, affect temperature. On the other hand, imposing only the gas constraint shifts the uniform distribution of HS towards the left. However, when both gas and temperature constraints are considered, only higher HS values turn out to be non-implausible. The explanation of this is as follows: third parameters are present whose values are in fact constrained by temperature, and which are correlated with HS. Forcing these parameters to have temperature-compatible values rules out low values of HS.

Something analogous happens with the boiler efficiency parameter. Finally, the infiltration rate (which plays an important role in limiting temper-

ature oscillation in the building) is indeed mainly driven by the temperature constraint. The additional imposition of the gas constraint only slightly reduces its non-implausible range.

6. Conclusion and Extensions of the Work

This work has discussed robust ways of accounting for uncertainty in pre-calibration and calibration of building energy models. The methodology has been illustrated on an actual dwelling and its energy model, with the aim of simultaneously matching observations of different nature: energy (gas and electricity use) and environmental (temperature in two zones).

Our methodology operates in a framework that allows to account for different sources of quantified uncertainty. We have discussed typical sources, such as model discrepancy and measurement error, and shown that their magnitude can play a decisive role in assessing the goodness of a match between simulations and observations. We have therefore illustrated History Matching (HM), a statistical procedure that sequentially rules out regions of the input space where simulations cannot match observations.

To explore the model's response throughout the input space, HM makes use of statistical emulators. Once built and validated, an emulator only requires milliseconds on a personal laptop to predict the model response on a new input configuration, and its prediction uncertainty can be naturally embedded in the HM framework. The hmer R package allows to implement emulation and the whole HM procedure [34].

While HM generally proceeds in multiple waves, in the example study discussed in this work one wave was enough to identify the final non-implausible region. This is a peculiarity of this example, where remarkably precise emulators for all quantities of interest could be built over the whole input space. In general, one of the strengths of HM is that the above is not at all required. Some of the outputs may not even be emulated in early stages. The sequential procedure allows the researcher to add constraints as the procedure moves to later waves. A quantity that was difficult to emulate in a given wave often behaves more smoothly within the more constrained NROY region of a later wave, and it can therefore be emulated with great precision at that stage.

Also notice that in our case, the speed of the E+ model allowed us to run $1,000$ simulations initially. Such a large number of runs is, once again, not needed in general. In fact, HM only needs a few runs at the start. The procedure itself will guide the researcher to focus new simulations within

a sequence of smaller regions, that eventually converge to the final non-implausible region. In additiona experiments that has not been detailed in this work , we have randomly selected 100 of the original $1,000$ simulations and used these to train (80 runs) and validate (20 runs) the new emulators. However, the quadratic behaviour of the simulator was easily captured by the 80 simulations, yielding results that, after only one wave of HM, were practically identical to the ones shown in Section 5, where all $1,000$ simulations were used.

Whilst emulators are classically used as model surrogates in HM, it is worth observing that less precise surrogates of the model may also be used (*e.g.*, linear regression), especially during the first explorative stages of research. Using linear regression, for example, would only take a few minutes of computational power but still allows the researcher to get a feeling of where the non-implausible region lies, while operating in a sound uncertainty framework.

In terms of computational power, the efficiency of emulation over the original simulator is apparent. All emulation and HM computations carried out in this work were performed on a personal laptop with 16GB RAM and 1.9GHz processor. For gas, and without any parallelisation, the computations (emulation and HM at 10^7 inputs, for nine months) were carried out in less than 28 hours. While still several orders of magnitude faster than the model running times, this is relatively slow for emulation standards, due to the high number of training points used (700). With only 80 training points, the same computations on the same device are executed in 2h and 5min. If we were to use the original E+ simulator directly, 10^7 simulations would have instead required about 18 years (as our original $1,000$ simulations were performed in about 16 hrs).

Finally, extensions of the methodology illustrated here can be manifold. Suppose a simulator is exceptionally slow to run, and faster versions produce lower-resolution results. A Bayesian framework where accurate emulators of the expensive model can be built, by using results of the faster version as prior information, is illustrated in [40]. This approach is referred to as multi-level (or multi-fidelity) emulation; an application for transport of volcanic ashes is discussed in [41]. Applications in the energy sector may concern the case of multi-vector energy modelling across districts or even national level, which is becoming crucial as a diverse mix of renewable inputs are used to replace fossil-based alternatives. Another similar example is building fabric energy modelling at building, cluster or even urban level whereby increasing

the physical boundary and resolution of the model imposes a very heavy computational duty.

An additional extension of the methodology we have discussed concerns the HM of time series data. This requires the use of emulators for time-evolving systems, as discussed for example in [42]. Multivariate implausibility measures ([14, 43]) may then be used to assess the distance between predicted and observed time series. This however requires a thoughtful specification of the time-correlation not only for the time series in question, but also for MD and OE. A discussion of this is beyond the scope of this work, but the interested researcher may refer to [ref] for an example of application to ***.(Michael, Hailiang, suggestion here?)

References

[1] S. Imam, D. A. Coley, I. Walker, The building performance gap: Are modellers literate?, Building Services Engineering Research and Technology 38 (3) (2017) 351–375.

[2] M. Royapoor, A. Antony, T. Roskilly, A review of building climate and plant controls, and a survey of industry perspectives, Energy and Buildings 158 (2018) 453–465.

[3] ASHRAE, Guideline 14-2002: Measurement of Energy and Demand Savings, ASHRAE, Atlanta (2002).

[4] M. Royapoor, T. Roskilly, Building model calibration using energy and environmental data, Energy and buildings 94 (2015) 109–120.

[5] D. Hou, I. Hassan, L. Wang, Review on building energy model calibration by Bayesian inference, Renewable and Sustainable Energy Reviews 143 (2021) 110930. doi:10.1016/j.rser.2021.110930.

[6] M. Goldstein, N. Huntley, Bayes Linear Emulation, History Matching, and Forecasting for Complex Computer Simulators, in: R. Ghanem, D. Higdon, H. Owhadi (Eds.), Handbook of Uncertainty Quantification, Springer International Publishing, 2017, pp. 9–32. doi:10.1007/978-3-319-12385-1_14.

[7] M. C. Kennedy, A. O'Hagan, Bayesian Calibration of Computer Models, Journal of the Royal Statistical Society: Series B 63 (3) (2001) 425–464.

[8] A. O'Hagan, Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering & System Safety 91 (10-11) (2006) 1290–1300.

[9] P. S. Craig, M. Goldstein, J. C. Rougier, A. H. Seheult, Bayesian Forecasting for Complex Systems Using Computer Simulators, Journal of the American Statistical Association 96 (454) (2001) 717–729.

[10] T. L. Edwards, S. Nowicki, B. Marzeion, R. Hock, H. Goelzer, H. Seroussi, et al., Projected land ice contributions to twenty-first-century sea level rise, Nature 593 (7857) (2021) 74–82.

[11] A. L. Wilson, C. J. Dent, M. Goldstein, Quantifying uncertainty in wholesale electricity price projections using Bayesian emulation of a generation investment model, Sustainable Energy, Grids and Networks 13 (2018) 42–55.

[12] I. Andrianakis, I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, R. G. White, Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda, PLoS computational biology 11 (1) (2015) e1003968.

[13] P. S. Craig, M. Goldstein, A. H. Seheult, J. A. Smith, Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments, in: Case Studies in Bayesian Statistics, Vol. 121 of Lecture notes in statistics, Springer, 1997, pp. 37–93. doi:10.1007/978-1-4612-2290-3_2.

[14] I. Vernon, M. Goldstein, R. G. Bower, Galaxy formation: a Bayesian uncertainty analysis, Bayesian analysis 5 (4) (2010) 619–669. doi:10.1214/10-BA524.

[15] D. Domingo, I. Malmierca-Vallet, L. Sime, J. Voss, E. Capron, Using ice cores and Gaussian process emulation to recover changes in the Greenland ice sheet during the last interglacial, Journal of Geophysical Research: Earth Surface 125 (5) (2020). doi:10.1029/2019JF005237.

[16] I. Vernon, J. Liu, M. Goldstein, J. Rowe, J. Topping, K. Lindsey, Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions, BMC systems biology 12 (1) (2018) 1–29.

[17] D. Daly, P. Cooper, Z. Ma, Understanding the risks and uncertainties introduced by common assumptions in energy simulations for Australian commercial buildings, Energy and Buildings 75 (2014) 382–393. doi:10.1016/j.enbuild.2014.02.028.

[18] A. Rasooli, L. Itard, C. I. Ferreira, A response factor-based method for the rapid in-situ determination of wall's thermal resistance in existing buildings, Energy and Buildings 119 (2016) 51–61. doi:10.1016/j.enbuild.2016.03.009.

[19] A.-H. Deconinck, S. Roels, Comparison of characterisation methods determining the thermal resistance of building components from on-site measurements, Energy and Buildings 130 (2016) 309–320. doi:10.1016/j.enbuild.2016.08.061.

[20] X. Meng, B. Yan, Y. Gao, J. Wang, W. Zhang, E. Long, Factors affecting the in situ measurement accuracy of the wall heat transfer coefficient using the heat flow meter method, Energy and Buildings 86 (2015) 754–765. doi:10.1016/j.enbuild.2014.11.005.

[21] G. Ficco, F. Iannetta, E. Ianniello, F. R. d. Alfano, M. Dell'Isola, U-value in situ measurement for energy diagnosis of existing buildings, Energy and Buildings 104 (2015) 108–121. doi:10.1016/j.enbuild.2015.06.071.

[22] K. Gaspar, M. Casals, M. Gangolells, A comparison of standardized calculation methods for in situ measurements of façades U-value, Energy and Buildings 130 (2016) 592–599. doi:10.1016/j.enbuild.2016.08.072.

[23] G. Desogus, S. Mura, R. Ricciu, Comparing different approaches to in situ measurement of building components thermal resistance, Energy and Buildings 43 (10) (2011) 2613–2620. doi:10.1016/j.enbuild.2011.05.025.

[24] C. Hoffmann, A. Geissler, The prebound-effect in detail: real indoor temperatures in basements and measured versus calculated U-values, Energy Procedia 122 (2017) 32–37. doi:10.1016/j.egypro.2017.07.301.

[25] A. Marshall, R. Fitton, W. Swan, D. Farmer, D. Johnston, M. Benjaber, Y. Ji, Domestic building fabric performance: Closing the gap between the in situ measured and modelled performance, Energy and Buildings 150 (2017) 307–317. doi:10.1016/j.enbuild.2017.06.028.

[26] P. Baker, U-values and traditional buildings, Historic Scotland Conservation Group: Glasgow, UK (2011).

[27] H. Yassaghi, N. Mostafavi, S. Hoque, Evaluation of current and future hourly weather data intended for building designs: A Philadelphia case study, Energy and Buildings 199 (2019) 491–511. doi:10.1016/j.enbuild.2019.07.016.

[28] T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, S. P. Corgnati, Advances in research and applications of energy-related occupant behavior in buildings, Energy and buildings 116 (2016) 694–702. doi:10.1016/j.enbuild.2015.11.052.

[29] W. Tian, Y. Heo, P. De Wilde, Z. Li, D. Yan, C. S. Park, X. Feng, G. Augenbroe, A review of uncertainty analysis in building energy assessment, Renewable and Sustainable Energy Reviews 93 (2018) 285–301. doi:10.1016/j.rser.2018.05.029.

[30] S. Coveney, R. H. Clayton, Fitting two human atrial cell models to experimental data using Bayesian history matching, Progress in biophysics and molecular biology 139 (2018) 43–58. doi:10.1016/j.pbiomolbio.2018.08.001.

[31] M. Goldstein, D. Wooff, Bayes Linear Statistics: Theory and Methods, Wiley Series in Probability and Statistics, John Wiley & Sons, 2007.

[32] C. E. Rasmussen, C. K. Williams, Gaussian Processes for Machine Learning, The MIT Press, Massachusetts Institute of Technology, 2006.

[33] F. Pukelsheim, The three sigma rule, The American Statistician 48 (2) (1994) 88–91. doi:10.2307/2684253.

[34] A. Iskauskas, hmer: History Matching and Emulation Package (2022). URL https://CRAN.R-project.org/package=hmer

[35] J. Kragh, J. Rose, H. N. Knudsen, O. M. Jensen, Possible explanations for the gap between calculated and measured energy consumption of new houses, Energy Procedia 132 (2017) 69–74. doi:10.1016/j.egypro.2017.09.638.

[36] J. L. Loeppky, J. Sacks, W. J. Welch, Choosing the Sample Size of a Computer Experiment: A Practical Guide, Technometrics 51 (4) (2009) 366–376. doi:10.1198/TECH.2009.08040.

[37] N. S. Lord, M. Crucifix, D. J. Lunt, M. C. Thorne, N. Bounceur, H. Dowsett, C. L. O'Brien, A. Ridgwell, Emulation of long-term changes in global climate: application to the late Pliocene and future, Climate of the Past 13 (11) (2017) 1539–1571. doi:10.5194/cp-13-1539-2017.

[38] C. A. Pope, J. P. Gosling, S. Barber, J. S. Johnson, T. Yamaguchi, G. Feingold, P. G. Blackwell, Gaussian Process Modeling of Heterogeneity and Discontinuities Using Voronoi Tessellations, Technometrics 63 (1) (2021) 53–63. doi:10.1080/00401706.2019.1692696.

[39] M. D. Mckay, R. J. Beckman, W. J. Conover, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, Technometrics 42 (1) (2000) 55–61. doi:10.1080/00401706.2000.10485979.

[40] J. A. Cumming, M. Goldstein, Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations, Technometrics 51 (4) (2009) 377–388. doi:10.1198/TECH.2009.08015.

[41] N. J. Harvey, N. Huntley, H. F. Dacre, M. Goldstein, D. Thomson, H. Webster, Multi-level emulation of a volcanic ash transport and dispersion model to quantify sensitivity to uncertain parameters, Natural hazards and earth system sciences 18 (1) (2018) 41–63. doi:10.5194/nhess-18-41-2018.

[42] S. Conti, J. P. Gosling, J. E. Oakley, A. O'Hagan, Gaussian process emulation of dynamic computer codes, Biometrika 96 (3) (2009) 663–676. doi:10.1093/biomet/asp028.

[43] J. A. Cumming, M. Goldstein, Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments, in: A. O'Hagan,

M. West (Eds.), The Oxford Handbook of Applied Bayesian Analysis, Oxford University Press Oxford, 2010, pp. 241–270.