A Bayesian Emulator Methodology to Support Evidence-Based Building Energy Model Parameterisation and Uncertainty Analysis

Mohammad Royapoora, Dario Domingob, Hailiang Dub, Aaron Boranianc, Sara Walkera, Michael Goldsteinb

[a] School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU

[b] Department of Mathematical Sciences, Durham University, Durham, DH1 3LE

[c] Big Ladder Software LLC, Denver, CO 80202

**How many CV(RMSE) and MBE are in each energy subcat that qualify as calibrated model. Could it be that we chose a smaller parameter space that the emulator did not have enough information to be able to completely define the uncertainty space. Run and emulate over a bigger parameter space:**

**However, as the number of spaces increase the**

**Cut out the parameters that are not efficient in enabling expensive models to history match successfully,**

# Abstract

This work uses a systematic, evidence-based approach to parameterise a building energy model to be calibrated against monthly gas and electricity data and hourly temperatures. The emphasis is to avoid tuning the model is an ad-hoc manner and retain input parameters within the bounds of actual energy and temperature data, building thermophysical information and observed occupant activities. Expert judgement and product specifications are used to introduce uncertainty bands to train a Bayesian emulator in two annual waves of simulation each involving 1000 runs.

The results show:

Having put the deterministic occupancy and operation of the building as closely as possible, and having used local weather data, the probable variation in fabric properties and plant efficiencies could only bring 7 out of 12 months into a calibration band of +\_15% (CV(RMSE)), with the rest of uncertainties having to be left to micro-climate, weather, stochastic occupant behaviour, ???

# Introduction

Stating uncertainty bands and integrating them in building performance simulation results is essential for producing high-quality results that acknowledge modelling limitations. Both model uncertainty and performance gap between model predictions and actual operational data remain areas of active building research. A recent article reported that a large proportion of building modelling community lacks essential knowledge on what the most fundamental parameter inputs for buildings are and how these impact model predictions [1]. This gap has major consequences, as retrofit strategies are mostly derived in consultation with building energy models and as such techno-economic benefits of proposed solutions can be misleading. Similarly, future buildings are expected to be more responsive to other civic activities (i.e. power generation and storage, transport, etc.)[2]. This can only be assessed by modelling, simulation and near-real-time analytics of a cluster of buildings at district and potentially city level which in turn requires generating accurate building energy data.

A preliminary step to support model-based decisions is model calibration. Calibration is the process though 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, evaluating in each case one or more measures of the discrepancy between the corresponding output and the observed data. In the context of building energy consumption, ASHRAE guidelines [28] are often followed *[refs]* to identify 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 (or input space) grows exponentially with the latter’s dimension. To meet the computational challenges involved, a new area of Bayesian statistics has developed, known as emulation [43, 54-56].

An emulator is a statistical surrogate of the simulator, providing predictions of the simulator’s output at inputs which the original model has never been evaluated at. The emulator can be built and validated on a personal device. It is instantaneous to run and therefore allows to predict the model’s prediction at 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 [58] and energy price [60] projections, epidemiology [59], hydrocarbon reservoirs [46], galaxy formation [44].

Emulation is developed under Bayesian principles, arguably the most natural to deal with uncertainties. In the context of calibration, key uncertainties that are often overlooked, but may significantly affect results, are the ability of the model to reproduce the process for which observations are available, and the intrinsic accuracy of such observations. To deal coherently with these and potentially other sources of uncertainty, we propose here the use of so-called implausibility measures to compare model (or emulated) outputs and observations.

These have been employed in the contexts of ice sheet shapes [47], gene evaluation [45], galaxy formation [44] and more, within a process known as “history matching”. The process sequentially removes regions of the input space deemed implausible to produce outputs matching 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 introduced and illustrated within an example study of a model predicting energy consumption in a single-house building. A link to an R package is provided, which helps the interested researcher to implement the proposed methodology within their own research.

This paper is organised as follows. In Section 2 the problem of inference from computer models under uncertainty is discussed. Limitations of current approaches are highlighted and the concept of model discrepancy is introduced. The methodology we propose is discussed in Section 3. Section 4 gives details on the energy model used as example study in this work and on the data used to calibrate it. Results are discussed in Section 5, with quantification of uncertainty reduction obtained by using an emulator rather than a linear model, and by carrying out calibration potentially in more waves. Discussion and conclusions are in Section 6.

# Uncertainty Quantification of Computer Models

## Calibration under Uncertainty

Calibration of computer models (*a.k.a.* simulators) is a crucial topic not only in the area of energy models, but in a variety of scientific disciplines. It consists in identifying values of the parameters of a computer model whose corresponding outputs match observed data. Mathematically, the model can be seen as a function, taking as input the list of model parameters of interest, and returning as output. Denoting by the real-world observations of the process being simulated, the calibration task can be stated as finding the inputs so that is “close enough” to .

Proximity between the two quantities 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 *[refs to some of Michael’s papers]*, two sources of uncertainty usually play a prominent role in the calibration of building energy models and many others.

1. *Measurement Error*. Energy consumption observations, as much as any observation, suffer from measurement errors. Their magnitude depends on the employed measuring device.
2. *Model Discrepancy.* Modelling assumptions and numerical approximations, albeit unavoidable, make a model an imperfect representation of reality. The simulated output will be different from the (average) value that would be observed in the real world under the same conditions specified in. We refer to this difference as model discrepancy, which we discuss in more detail for energy models in Section 2.3.

Accounting for the above sources of uncertainty is crucial when assessing whether a model has been successfully calibrated against observed data.

## Limitations of Current Approaches in Energy Building Models

In the context of building energy models, ASHRAE guidelines [28] are often followed to assess whether a model has been successfully calibrated against observed data [18,42]. The procedure consists in computing the following discrepancy measures between a sequence of simulated model outputs,, and a sequence of observations, , :

(1)

. (2)

The model is then considered calibrated if the relevant one of the two following conditions is met:

1. -10% ≤ MBE ≤ 10% and CV(RMSE) ≤ 30% for hourly measurements.
2. -5% ≤ MBE ≤ 5% and CV(RMSE) ≤ 15% for monthly measurements.

While the above criteria are easy to check, their use to assess model calibration presents some limitations. For example, the provided thresholds are independent of both the level of accuracy to which measurements are available, and the level of discrepancy between the model and reality. Moreover, formulas (1) and (2) can only be applied to some of the physical quantities for which model outputs and observations may be available.

Indeed, the two formulas provide a measure of how average bias and average magnitude of the sequence compare to the average measured value. For a quantity such as energy consumption, and in fact for any quantity on an intrinsically positive scale, this is meaningful: the two formulas provide a clearly defined, dimensionless value. It is however less so for other quantities, such as temperature. Here, the use of different units would lead to apparently different results. In reality, only Kelvin should be used to ensure positivity, but criteria a) and b) would then become of little help, essentially always fulfilled due to the large value of the denominator in (1) and (2).

The approach to calibration that we propose in Section 3 overcomes the previous issues. The statistical setting in which it arises makes it easy to account for recognised uncertainties during the process. It is applicable to any quantity of interest and, thanks to the use of emulation, allows to explore the whole space of interest by assessing compatibility between outputs and observations at millions/billions of inputs, in little time.

## **Model Discrepancy in Building Energy Systems**

Multiple factors drive energy consumption in a building, such as quality of wall insulation, external weather, occupants’ lifestyle, etc. While current models such as Energy+ attempt to reproduce the building’s physics of with high fidelity, no model can include an exact representation of all the features affecting environmental properties (*e.g.* internal temperature) and energy consumption in a building. Simplifications are necessary to keep the model complexity to reasonable levels. Example of factors which are commonly either neglected or included via simpler dynamics are as follows:

1. The dynamic nature of fabric thermo-physical properties: most models consider the U-value of masonry walls as constant, although its value changes as a function of wall moisture content. *(Hygroscopic material can indeed exhibit wide thermophysical variance at different moisture and temperature conditions.)*
2. The occupant’s behaviour and interaction with the building (window opening, light and small power usage): their stochastic nature is rarely accounted for in energy models.
3. Uncertainties and variations in plant operational characteristics: these are extensively simplified in energy models.
4. Exact zone air exchange figures and fabric infiltration values: they are difficult to determine, and fabric infiltration values can only be estimated after performing a building pressure test that is logistically difficult and costly, particularly for larger occupied buildings.
5. The available weather files: that impose large uncertainties in particular with solar irradiance data that quite often is partially or fully modelled (as opposed to measured). Micro-climatic variations are widely ignored and understudied. An example could be the difference in conventional airport weather station data used to model urban settings. Airports are exposed terrains often with proximity to water to facilitate emergency aircraft landing. Annual weather files compiled in these locations would therefore report higher wind velocities and miss the heat-island effect that is increasingly separating urban and open country micro-climates.

None of the above assumptions and approximations undermines the validity of the model. The latter remains a key tool for the researcher to gain insight on the real-world phenomenon under study, such as energy consumption in a building. However, awareness of similar assumptions and approximations is key to quantify the discrepancy thereby induced between model outputs and real-world observations.

# Methodology

The aim of this section is to lay out a statistically robust procedure to deal with uncertainties when calibrating a model. The flowchart in Figure 1 illustrates the proposed methodology. This aims to sequentially rule out regions of the model’s parameter space where a match between model outputs and observations can be confidently rejected given the involved uncertainties. This process is often referred to as history matching, the name originally indicating to the attempt to “replicate history” within a study on hydrocarbon reservoirs [46].

The procedure makes use of two tools to accomplish the task:

1. Emulators: to instantaneously predict the model output at parameters where the model has not been run, and quantify the prediction error;
2. Implausibility measures: to quantify the distance between model outputs and observations, in light of different sources of uncertainties.

The first two steps of the methodology (blue in Figure 1) consist in quantifying the uncertainties in the model and in the observed data (energy consumption for the sake of this work, but potentially any observable quantity), and in training an emulator of the model in question. The two steps are independent of each other, and are discussed in Section 3.1 and 3.2 respectively. The following steps (yellow in flowchart) require to evaluate the implausibility measure at a very large number of inputs over the space where matches are being sought, so that regions with highly implausible matches can be ruled out. At this point, the researcher may wish to repeat emulator training and implausibility evaluation on the Not-Ruled-Out-Yet (NROY) part of the space: the simulator often displays simpler, more linear behaviour locally, hence a new emulator over the smaller region can be more precise and allow the researcher to exclude further regions (red). Sections 3.3 details the definition of implausibility measures, while their use for the above tasks is discussed in Section 3.4. Finally, within the region deemed non-implausible to yield a match, we propose a way to sample points, at which the actual simulator can be run to check that a match has been achieved.

## Quantifying Model Discrepancy and Measurement Errors

Figure . Flow chart illustrating the methodology proposed in this study.

The concepts of model discrepancy (MD) and observational error (OE) are not unique to the context of energy model calibration, they are relevant to any uncertainty analysis linking computer models and measurements. In fact, the structure we present here has been successfully applied to a wide range of contexts [44-47]. Within the context relevant to this work, the reader may think of the computer model as predicting energy consumption in a building. The prediction depends on the values taken by some model parameters, such as the ones in Table 1. We call the vector of parameter values an input to the model, . The corresponding modelled consumption is .

Following [43], we assume that an appropriate choice of inputs exists, , that accurately represents the values of the system’s parameters. The real-world value of the modelled quantity and the simulator output at the above choice of inputs are then linked via the formula:

where the model-discrepancy term accounts for the difference between real and simulated process.

The (unobservable) real-system value is usually estimated through measurements of the process of interest. We link these two quantities via the relationship:

where the term accounts for the observational error in the measurements.

The additive formulation in equations (1) and (2) is a simple but efficient way to model MD and OE, which also makes statistical inference tractable. Information about the quantities and should in fact be sought in statistical form, rather than be 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 slightly 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 the discrepancy between the model at the “best” input and reality. In the example study discussed in this paper (Section 4) we consider both 10% and 20% MD and discuss how the choice affects calibration results. For a more detailed treatment of MD, and the further distinction between internal and external MD, the reader is referred to [43].

## Bayes Linear Emulators as Fast Model Surrogates

When performing calibration, evaluating the model at different inputs is key to compare its outputs to observations. However, evaluating the model at a set of inputs that cover the parameter space thoroughly proves, practically always, time-wise and computationally unfeasible. Emulators may be employed to overcome the issue.

An emulator is a statistical surrogate of the model, predicting the model’s output at inputs where the latter has never been run. An emulator’s most evident advantages are its speed and low computability requirements: predictions can be obtained within milliseconds on a common laptop. Moreover, from a statistical perspective, the uncertainty statements accompanying the emulators’ predictions lay the basis for an uncertainty analysis of the model.

In this work we employ Bayes linear emulators (BLEs) to replicate the behaviour of the building energy model under consideration (described in Section 4). In this section, we give a brief overview of the choices behind the construction of such an emulator, referring the reader to [43-44] for further details. The interested reader is also referred to [61] for a much more in-depth treatment of Bayes Linear principles.

We assume the following form for, the output of the computer model at input:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Equation (1) is made of two terms. The first one is a regression component: it uses a linear combination of known functions to model the global behaviour of across the input space. The second term is instead a local term, meant to model the residual variability in once the regression component has been accounted for.

Suppose the model has been run at a sequence of inputs. We call each a design point, and will denote the corresponding output by

* The coefficients in equation (1) can be estimated by fitting a linear regression model with predictors to the known pairs.
* Values of the residual process will then be known at each design point,. These values will approximately oscillate around 0, with patterns that the regression term has not been able to detect. We predict values of via a BLE.

The idea is to consider as a stochastic process, specifying a prior mean and covariance structure for it. In the Bayes linear framework, these can then be adjusted in light of the observed values at. This provides a prediction (the adjusted mean) and an uncertainty statement (the adjusted variance) for the value, at any.

The prior specifications we make are as follows. We assume to have mean zero at all inputs. As to the prior covariance between the values of at any two inputs and, in this work we use the squared exponential covariance function. The further the inputs andare, the lower this is:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

The subscript in and denotes the component of the two inputs, *i.e.*, identifies one of the model parameters. The positive coefficient (called correlation length, or length scale) measures the strength of correlation in outputs when the parameter is modified. Finally, the coefficient denotes the prior variance of at all inputs.

Expression (2) has been written in terms of all components of the input vector. However, just some of the parameters often prove relevant to explain essentially all the variability in. We call these active parameters and, in practice, only use them in expression (2). The small variability left in, due to the inactive parameters, is modelled as uncorrelated noise with constant variance.

Notice that no particular distribution has been assumed for the process. Within the Bayes linear framework, we only need to specify mean and covariance, as above. These are then adjusted to the values, yielding an adjusted mean and variance for the value of at any.

Although, strictly speaking, the above procedure builds a BLE for the “residual” process, when the regression prediction is added to the mean of the BLE we get an overall prediction for. Ultimately, this is what we are interest in. For this reason, unless otherwise stated, in the following the term emulator will refer to the overall emulator of. At any input, this provides a prediction of the unknown value, and a standard deviation quantifying the uncertainty of such prediction.

We have not provided details of the adjusted mean and covariance formulas here, these can be found for example in [43], [44], [61]. 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, notice that these are: the predictors of the regression term; the variances and; the active parameters; the correlation lengths. An R package to perform Bayes Linear emulation, used to build the emulators in this work, can be downloaded at [put link].

## Implausibility Measure and History Matching (calibration)

At any input, the emulator provides a prediction of the model output and an associated standard deviation. This section discusses the use of this information to assess how implausible the input is to match the real-world features having led to observations. The methodology is applicable to a variety of contexts. For the sake of illustration, we make reference to the case where the model predicts energy consumption in a building as function of the building’s parameters (the real-world features to be matched), and is the measured consumption in the building.

The choice of measuring implausibility as opposed to plausibility is a recognition of the uncertainties involved, as clarified below. Indeed, for a given, we would like to compare the two following quantities:

1. The consumed energy in the building, if the building parameters were equal to.
2. The actual consumption in the building.

None of the two quantities, however, is directly available. The first one can be estimated by, the model prediction at. The second one is approximated by a measurement, usually a meter reading. Model discrepancy (MD) and observational error (OE) make and not the same as the quantities in i) and ii). In the notation of section 3.1, these are instead and, respectively.

Thus, we cannot certainly establish whether the quantities in i) and ii) are the same. But, through the difference and given the tolerance allowed by MD and OE, we can establish whether the quantities in i) and ii) *may* be the same (inputnon-implausible) or not (inputimplausible). In other words, we are interested in the following quantity:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

The higher its value, the more “implausible” the input will be.

Notice that expression (3) requires knowledge of the simulated value. In practice, this is not available for most inputs. Instead, we can use the (easily available) emulator’s prediction, and define:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

The addition of the emulator variance in the denominator accounts for having replaced the unknown model output with its emulated prediction. Expression (4) will be referred to as the implausibility of input. By using the known quantities and as proxies for the unknown quantities in i) and ii), it provides a statement on the possibility that the latter two be the same. Higher values of make this more implausible.

## History Matching (calibration)

Implausibility measures are fast to compute, since they make use of emulator predictions in place of expensive model runs. They can therefore be used to explore thoroughly the input space, and discard inputs which are highly implausible to yield a match with the observations. In the context of UQ of computer models, this process is referred to as history matching (HM) [46]. We illustrate it below. Comments on the procedure and its advantages in handling uncertainties with respect to other procedures such as the one in Section 2.2 will follow.

History matching is usually carried out in waves, which aim to sequentially rule out implausible parts of the input space. The procedure is repeated till either the non-implausible region stays unchanged, or every part of the input space has been ruled out as implausible (we discuss inference for this case below). The transition between waves is made possible by performing additional model runs within the not-ruled-out-yet (NROY) region and training a new emulator in the region. Schematically, we summarise history matching in the following steps.

1. Choose a positive threshold: an input will be considered implausible if. We discuss and advise on appropriate values of in section 5.4.
2. Define the current NROY region as the one for which.
3. Sample additional input points in the current NROY region, and perform the corresponding model runs.
4. Use the model outputs to build a new emulator of the model in the region and compute the new implausibility measure, .
5. Identify the new NROY region (via) and compare to old one.
6. Repeat steps 3-5, till NROY region stays unchanged, or is empty.

At each wave, the percentage of space that is ruled out depends on several factors. However, particularly in earlier waves, it is not uncommon to rule out the majority of the current NROY space, especially when multiple constraints are considered (we discuss and illustrate this in Section 5). This usually makes the emulator in step 4 be considerably more accurate and less uncertain in the current NROY region, than its predecessor was. This is because the way the model responds to its inputs may vary significantly across the input space but be simple to emulate in a small region. Hence in a few waves, even one or two, the procedure usually leads to identify only a small fraction of the original input space where a match is likely.

We conclude this section with some notes. While HM can be performed on a laptop, it is usually more time consuming than other classical calibration procedures, such as the ones described in Section 2.2 where a number of model runs are compared to observations. However, there are a number of advantages that make HM a more robust and a more successful calibration procedure.

First, by using emulation and focussing model runs only where needed, HM allows a thorough exploration of the model input space, while keeping the overall number of runs low. This is key 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. If only an (even high) number of simulations are scattered among the input space, the subset of these which match observations is likely to be empty, but the region where a match is possible is in fact not.

Disadvantage: need to run runs more than once

* Uncertainties about model discrepancy and measurement error are accounted for when comparing model outputs and observations
* HM allows to explore the input space thoroughly, while only a relatively low number of runs are used. This is achieved by sequentially focussing new runs only when it is needed.
* In a high dimensional space, the non-implausible region may only represent a tiny fraction of the original space, and classical method may easily fail to detect it. HM will instead

. The methodology is allows to explore the input space thoroughly, while only requiring to run the model a relatively low number of times. Moreover, runs are increasingly concentrated in the region

# Example Study

This section describes the example study we use to illustrate the methodology outlined in Section 3. This is the one of a single house, of which a model has been created in EnergyPlus, and for which energy consumption and temperature data is collected to calibrate the model.

## 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 principally 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 range, using a predictor/corrector process. The uniqueness of E+ lies in being a physically based modelling solution that oversees a simultaneous calculation of radiative and convective heat and mass transfer processes, adsorption and desorption of moisture in building elements, iterative plant, building fabric and air interactions and accurate temperature and comfort predictions. This integrated and simultaneous simulation process is completed via several modules (and overseen by EnergyPlus 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 interestingly the zone air heat balance equation is one of the primary mechanisms that describes the connected nature of the loads within a zone, the corresponding plant duty to offset these loads as well as the zone mean air temperatures as follows:

[1]

Where is the rate of change of thermal energy stored in the zone air, is the sum of convective internal loads, is convective heat transfer from the zone surfaces, is the change of the room air enthalpy as a result of zone air mixing, is infiltration heat transfer and finally is the HVAC system input to achieve its target temperature. Given that E+ assumes a uniform zone air and surface temperature, uniform long and short-wave radiation and diffuse radiation and reflective surfaces (as opposed to direct or point-based), it is reasonable to regard zone air temperature as the interconnection where connective, radiative and convective heat balance and mass transfer are realised. Essentially each item on the right-hand side of Eq. 1 indicates a change of enthalpy due to environmental perturbations, while the left-hand side describes how these perturbations impact the zone air temperature and the enthalpy that it holds.

## Building

A detached two-storey masonry construction built in 1994 was selected as the case-study building. 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 activity. The building (with a gross area of 168.66m2 and 19.73m2 of unheated space) is in a built-up urban surrounding, is only partly shaded (on its west elevation) by another adjacent property (which was considered within the modelling work). Across the monitoring year (2016) the property had an observed annual gas (15,381 kWh) and electricity (2,991 kWh) consumptions that respectively reflect high and medium UK typical domestic consumption values [23]. Occupants only utilised shower facilities which at a measured flow rate of 4.37 l/min and recorded average eight 20-minute showers per week correspond to an average of 50 l/person/day. These recoded 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 their heavy use of gym washing facilities. Gas cookers (containing 3kW and 5kW hubs) were used 4 times a week for 1 hr per cooking session.

## Energy and temperature 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 sensors (two positioned in the south facing master bedroom and two in north facing kitchen). Whereas electricity and gas data required no imputation, overall kitchen and master bedroom temperature sensors had total annual losses of 5.7% and 2.7% that required imputation. Each missing hourly temperature cell was imputed by the average of the previous and successive available cells. Gas consumption data was manually recorded at monthly basis using mains gas meter. Electricity was logged at 10s intervals using two clip-on current sensors on the incoming live cable (to reduce measurement uncertainty) and the two sets of similar readings were averaged and aggregated to form the measured power usage.



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

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

## Input parameter range selection

By consulting manufacturers specification and the house builder’s literature, a detailed set of parameter inputs were compiled and where the greatest quantifiable uncertainty existed, lower and upper bands were imposed on the input value used. These bands were so far as possible derived from scientific literature and used to dictate the size of associated variations explored in batch-runs (Table 1). Ground floors are less prone to variations in internal and external air velocities that act on walls and roofs more robustly, and lead to dynamic heat transfer values that fail to be captured by standardised calculation methods. Therefore, a smaller floor uncertainty margin was derived from literature and imposed on floor thermal resistance (Tables 2 and 3).

The property’s glazing was updated in 2009 and manufacture’s literature set the G and U-value of the fenestration to 0.691 and 1.788 W/m2K respectively, with respective error bands of ±5% and ±2%. The compound upper and lower limit of these two values altered the gas consumption of the calibrated model by ±2.05 kWh (± 0.013%). Given its negligible nature, the error bands of the glazing were discounted in batch simulations. The compound effect of all other uncertainty bands created a lower boundary of 8,842 kWh and an upper boundary of 26,452 kWh with respect to an observed gas consumption of 15,381 kWh (i.e. -42.5% to +72%). Given that even a 5% increases in fabric U-value was reported to raise energy consumption of family homes by 0.3-2.5% [26], a uniformly distributed uncertainty band is imposed on elemental U-Values to reflect similar magnitude of variations reported in literature, as outlined in Table 2.

Actual building infiltration rates are difficult to arrive at and require convoluted air permeability tests. Table 4.16 of CIBSE guide A [6] outlines a range of 0.25 to 0.95 air change per hour (ACH) for various 2-storey buildings below 500m2 with a value of 0.5 ACH describing typical constructions similar to the case-study building. Therefore 0.5 ACH informed the calibrated base model with 0.25-0.95 ACH representing the range of possibilities that batch simulations explore (Tables 2-3). Local weather files compiled by a weather station approximately 3 miles away from the site was used to support the model development [27].

# Results

This section illustrates the application of the methodology we propose to the example study described in Section 4. We build emulators of some of the outputs of the building energy model under consideration, and use these to carry out a thorough exploration of the model’s input space in a reduced amount of time. Thus, within the statistical framework discussed in Section 3, we use the emulator predictions to identify subsets of the input space yielding a match with the observed data.

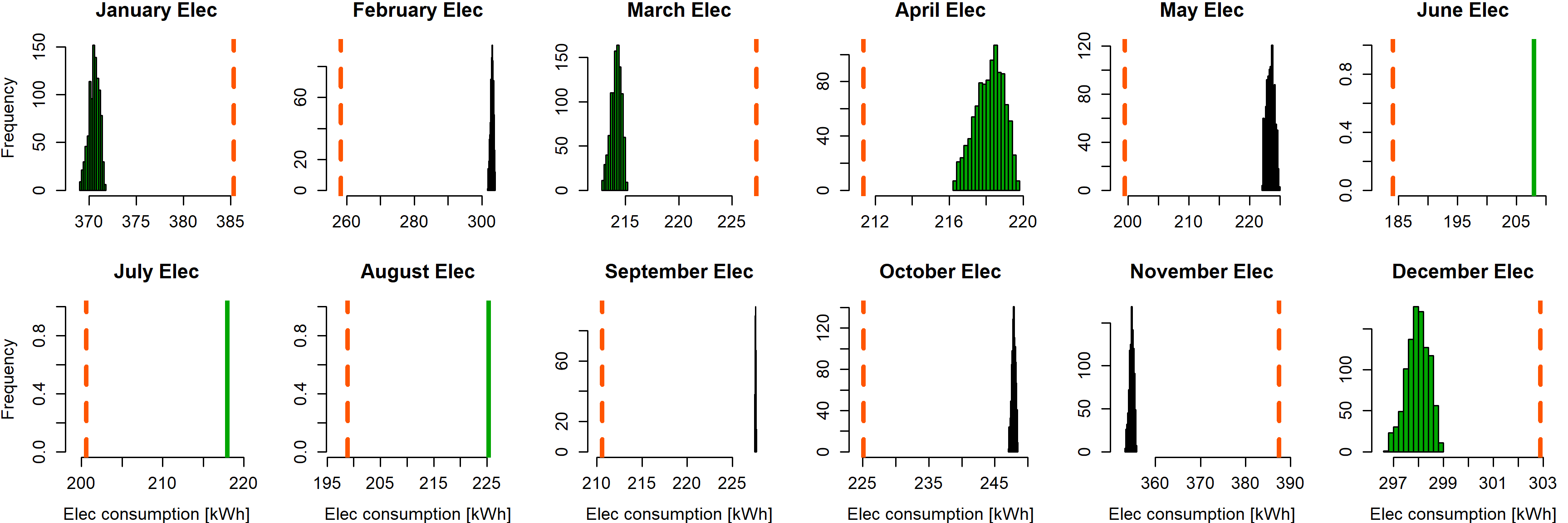
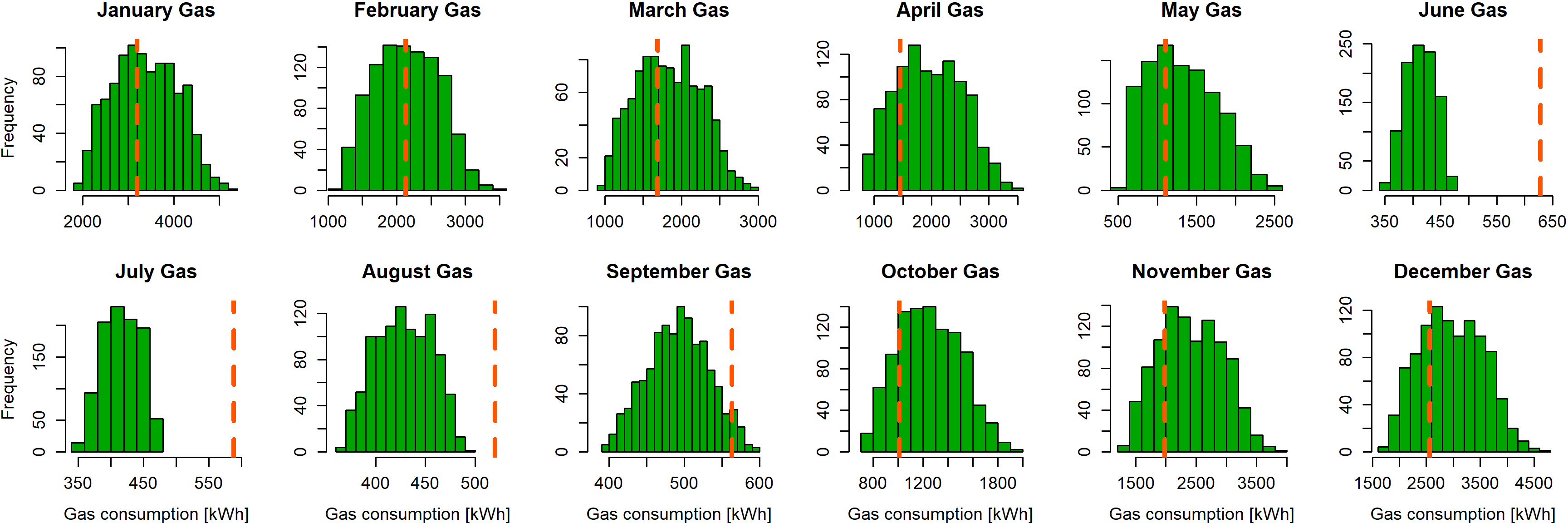
## Experimental Design and Simulated Results

Table 3 shows the eight variables varied in the model and, for each of them, the range over which a match with the data is sough.

In order to build emulators of the monthly energy consumptions, the outputs of a small number of model runs are needed. An informal rule of thumb suggests to use a number of runs about 10 times larger than the number of parameters varies (i.e., of the input space dimension) [48]. Note that a lower/higher number of runs may however prove sufficient/necessary to build an accurate emulator, according to the complexity of the model’s response to its inputs. In our case the simulator was reasonably fast to evaluate (*specify estimate of time*), and a total of *n*=1,000 simulations were run. This is substantially more than what typically needed for eight inputs. Throughout most of this section (up to and including 5.4), we harness the information provided by all 1,000 runs to build and train the emulators. However, in Section 5.5, we simulate the case where only 100 runs are available, discussing the gain obtained by later including additional 100 runs, within a small neighbourhood of the space classified as non-implausible after the first wave of emulation.

Figure . Distribution of simulated gas (panel a) and electricity (panel b) consumption, for each month, at the n=1000 inputs chosen in this study through LHS. The dashed vertical line in a plot denotes the observed consumption for that month. In the case of electricity, the three summer months (Jun-Aug) show no variation in simulated output among the n runs: the single simulated value is identified by the green solid line.

(a)



(b)

The experimental design, *i.e.* the set of *n* inputs at 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 *nxn* uniform grid of the square. The same idea generalises to higher dimensions. For this reason, LHS is commonly used in the design of computer experiments analysis, [44], [50], [51], to identify points in a hypercube that fill the space well and are not too close to each other [49]. LHS is easily performed in R via the “lhs” package.

### Gas

Panel (a) of Figure 2 shows simulated and observed values of monthly gas consumption. 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 points within the explored space (an 8-dimensional hypercube with side ranges as in Table 3) may yield 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 is to be ruled out 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, however, that in a real case study causes for the model’s underestimation of summer consumptions should be identified and acted upon, prior to calibration. If the model dynamics is reliable, one reason for such a mismatch is likely to reside in having fixed one or more model parameters to a wrong value in all simulations. In this case, values for this parameter should instead be varied in the experimental design and the parameter added as further input over which to perform calibration.

### Electricity

Panel (b) of Figure 2 shows observed and simulated electricity consumption over the same design. All months display little variability in output, with an observed consumption outside of the range of simulated consumptions. Moreover, the three summer months show no variation in simulated output among the *n* runs. [*add plausible reasons for this, eg important parameter not varied. What could it be?*]

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. Again, in a real case study, the first step to take in this case would instead be to detect which parameters affect the electricity consumption, and ensure that these are varied in the experimental design before calibration is performed.

## Emulation of Monthly Gas Consumption

For each of the nine non-summer months, the *n=1,000* simulations provide a dataset of n pairs (), where is the simulated gas consumption associated with input. We use the dataset to build an emulator of that month’s gas consumption. To that aim, each of the eight input variables is first linearly rescaled so that it ranges in the interval [-1,1].

Recall from Section 3.2 that various choices (*e.g.*, on covariates, correlation lengths, prior variance) have to be made when building an emulator. To allow to validate the emulators, 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 3.2, the prediction of the simulated consumption at a point is computed as sum of: i) a linear regression part; ii) a prediction of the regression residual at the point , in the form of a Bayes linear emulator. Details on each of the two parts follow.

### Linear Regression

The only choice to be made in building a linear regression model concerns the predictors to use. For all the months of interest, a preliminary exploration reveals that the response (gas consumption) is very well explained as quadratic function of (a subset of) the eight input variables, which in the following we call for convenience. Thus, we proceed as follows.

Let be the set of all mutually orthogonal linear, quadratic and interaction terms of .[[1]](#footnote-1) For a given integer , consider the linear model with highest coefficient of determination among all linear models with exactly of the covariates in . As it turns out, for all nine months, the improvement obtained by incrementing the number of covariates by one is statistically very significant ( at least up to . However, even 10 covariates give an adjusted- which is notably high, about 0.999 across all moths. To keep the approach uniform among months, we then consider the linear model with the “best” 10 regressors for each month.

Note that, in general, such a high should raise concerns of overfitting. In our case, however, the concern can be ruled out by observing that only 10 regressors have been used to explain 700 responses. The high coefficient of determination mirrors an intrinsically quadratic model dynamics.

### Emulators of the Residuals

To build an emulator of the regression residuals of each month’s consumption, choices about the following quantities in equation (\*\*) are to be made: active inputs, correlation lengths, prior emulator variance and nugget variance. We proceed as follows.

* **Active inputs** :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 inputs. 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** : Once the active inputs are chosen, the same correlation length *d* is used for all of themfor all relevant i and j. The value of *d* at each month is chosen by assessing the emulator's performance on the evaluation set, and is reported in Table 2. Note that, to attain a similar level of correlation across the space, higher correlation lengths are used when a higher number of active inputs is present.
* **Prior variances** **and**: The cumulative prior variance is set equal to the variance of the residuals being fitted. This overall variance is split into the two components, and , in the proportions of 95*%* and 5*%* respectively.

Table 2 provides details of all choices made to build each of the nine emulators, including the ones concerning the regression line. The function `Linear\_Bayes\_Emulation` within the R package [put link here] can be used to perform emulation. We discuss validation of the emulators in the next section.

### Emulator Validation and Performance

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Predictors** | **Adj.** |  | **Act. Inputs** |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Jan |  | 0.9998 | 85.94 |  | 0.8 |
| Feb |  | 0.9998 | 52.30 |  | 0.65 |
| Mar |  | 0.9997 | 50.14 |  | 1.3 |
| Apr |  | 0.9998 | 50.04 |  | 1 |
| May |  | 0.9989 | 206.45 |  | 1 |
| Sep |  | 0.9994 | 0.91 |  | 1.2 |
| Oct |  | 0.9996 | 24.40 |  | 1.4 |
| Nov |  | 0.9998 | 57.35 |  | 1.2 |
| Dec |  | 0.9998 | 68.77 |  | 1.3 |

The active inputs and correlation lengths in Table 2 are chosen based on the emulator’s performance on the evaluation set. The latter consists of 150 pairs not used in the emulator’s training, where each is the simulated output at input. At each such, the emulator provides a prediction of the simulated output and an uncertainty statement about the prediction in the form of a standard deviation . One way to assess the emulator’s performance at is thus to compute the number of standard deviations which separate the emulator prediction from the known simulated output:

( )

Table 3: For each month: variance of the regression residuals to which an 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** |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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%] CI | 31.0 | 28.0 | 9.3 | 10.0 | 68.9 | 0.20 | 3.8 | 12.8 | 12.7 |

We call the emulator standardised error at . As we make no distributional assumption on the emulator, we can appeal to Pukelsheim’s 3-sigma rule [52] to constrain expected values of : 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 about 95% or more of the to lie between -3 and 3.

For each of the nine emulators, we consider the plot of versus the predictions . Given a good emulator, such a plot should be characterised by an approximately random scattering of the points around the line , 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 inputs 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 in general yield more than 95% of less than 3. This is to prevent making choices tailored to the specific points used for validation. Once the parameters and are chosen, we compute the standardised errors in equation *(1)* 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. *Plots of vs for the evaluation and test sets are shown in Appendix.*

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 uncertainty in modelling the residuals. Quantification of the latter is provided in Table 3, 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 when compared to the range of simulated outputs (Figure 2).

## 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 . We use the implausibility measure described in Section 3.3 to compare the prediction with the observed consumption . The observational error () and model discrepancy () are set as follows.

* The OE is set to 5% of the observed value , in agreement with the manufacturer’s instructions of the employed devices.
* 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 3.1, 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 value is likely to represent a good estimate in many energy applications. If there are reasons to believe 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.

## Calibration (history matching)

At a configuration of inputs, we compute the implausibility measure for each month: call this for month *M*. The higher, the more unlikely input is to be matching the features yielding the observed consumption for that month. In order to associate an overall implausibility to, we consider the maximum over all implausibilities:

( )

Hence, we consider an input as non-implausible if

( )

The main reason for choosing a higher threshold than the more common *[refs]* is that condition (3) is verified when nine simultaneous conditions are. A higher threshold is therefore needed to keep the probability of incorrectly rejecting an input small. Indeed, if is the probability of misclassifying an input as implausible for a single month, the probability of this happening for at least one of months increases with : *e.g.*, under an assumption of independence.

An alternative approach may be to define as the second-highest value of over all months, while keeping a relatively low threshold in (3). However, if a particular month proves more challenging to match, then this approach risks to classify several points as non-implausible, although those points are highly unlikely to match the observed consumption for that month.

Under condition (3), the percentage of space classified as non-implausible is 0.30% when 10% MD is used, and 19.52% for 20% MD. These percentage have been estimated by computing the proportion of non-implausible points in a sample of size , generated via a Sobol sequence in the 8-dimensional unit hypercube. The error on these estimates, approximately equal to where is the estimated fraction, is therefore of the order of and in the two cases respectively. This makes both estimates very accurate.

***Computational note***: on a laptop with 16GB RAM and 1.9GHz processor, the computation above took overall about 28 hours. Note that, while still several orders of magnitude faster than the model running times, this is relatively slow due to the high number of training points used (700). With 100 training points, the same computations are executed in 2h and 20min only.

### Non-Implausible Region

The non-implausible region lives in an eight-dimensional space, one dimension per input variable considered. To identify variables which play a role in constraining the region, one can look at two-dimensional scatter plots of the non-implausible points, for all possible pairs of the 8 variables. In the 10% MD case, the plots (not shown) reveal that two variables are particularly significant: infiltration rate () and cooking energy consumption (). One additional variable (heating setpoint, ) also seem to play a role in identifying non-implausible points, but to a lesser extent than the first two.

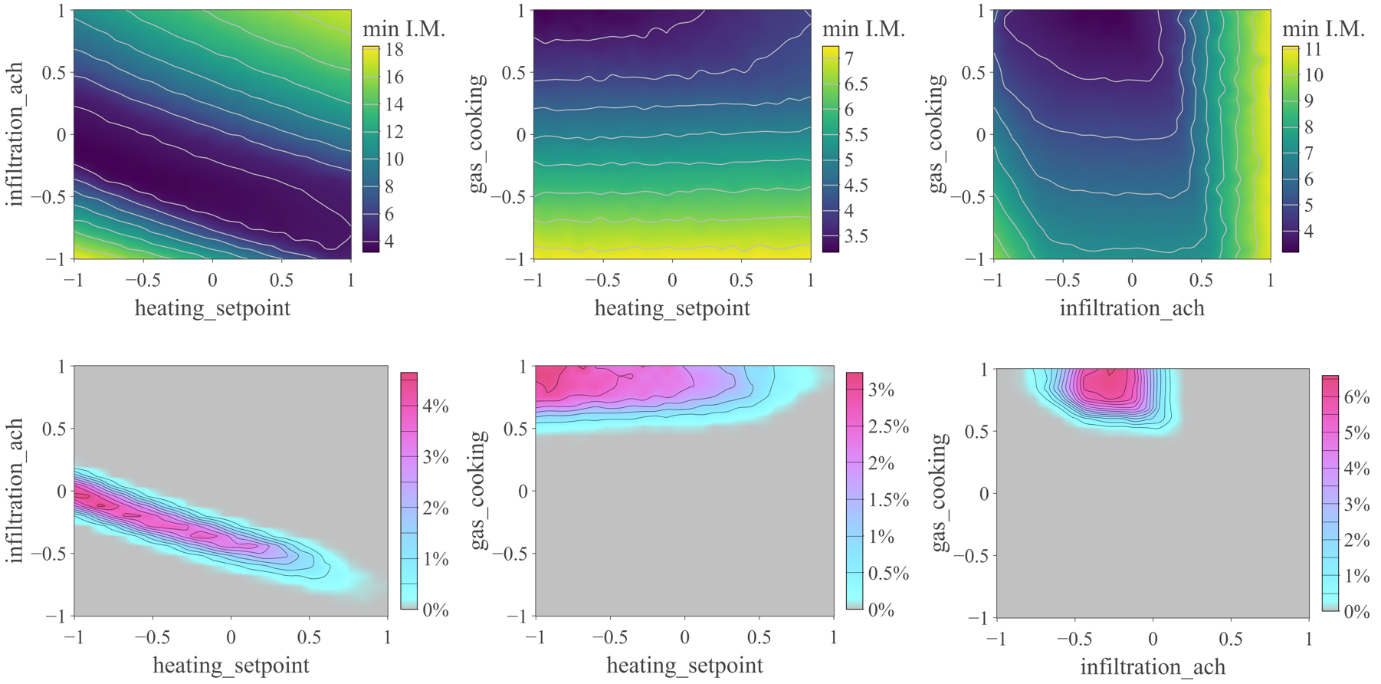
Figure 4 contains minimum-implausibility (MI) and optical-depth (OD) plots for all pairs of the three above variables. The MI plot of a pair of variables shows the minimum value of the implausibility measure over the hidden six dimensions. The OD plot instead computes which fraction of the hypervolume extending over the six hidden dimensions is non-implausible. Thus, values higher than 4 in a MI plot identify points that, irrespectively of values taken by the variables not represented in the plot, will always be classified as implausible. The associated OD plot instead provides information on the distribution of the remaining, non-implausible, points.

Figure . [**Note: I will make this figure nicer**: 3 by 3, with variable names and corresponding ranges on the diagonal, and no axis labels on the off-diagonal plots (which will also make these plots larger).] Plots concerning 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.

Figure 4 reveals that only values of the gas cooking higher than 5[unit here] 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 approximately between 0.25 and 0.65. Note that the two variables seem to be relatively independent. A stronger dependence can be seen instead for non-implausible values of heating setpoint and infiltration rate. Heating setpoint values are by themselves only minimally contrained: however, for a fixed infiltration rate, the range of non-implausible heating setpoint values is reduced.

The relevant subplots in Figure 4 suggest that non-implausible values of the heating setpoint may also be found to the left of the range originally deemed appropriate for this input variable. A similar consideration is valid for gas cooking values to the right of the relevant range. Similar findings are not uncommon, especially during a first wave of emulation. In such a case, the authors’ advice is to proceed as follows.

1. Use the available emulator(s) to explore the new parts of the space where a match is possible.
2. Compute the implausibility measure in the new region. Note: no runs of the original model are available in the region, which will generally lead to higher emulator uncertainty and therefore lower .
3. If non-implausible points are found in the region, run the original model on a number of them (according to computational resources).
4. Build a new emulator based on:
   1. Runs of the original design which fall into a suitable neighbourhood of the non-implausible region (*e.g.*, where ); and
   2. The new runs at point 3.

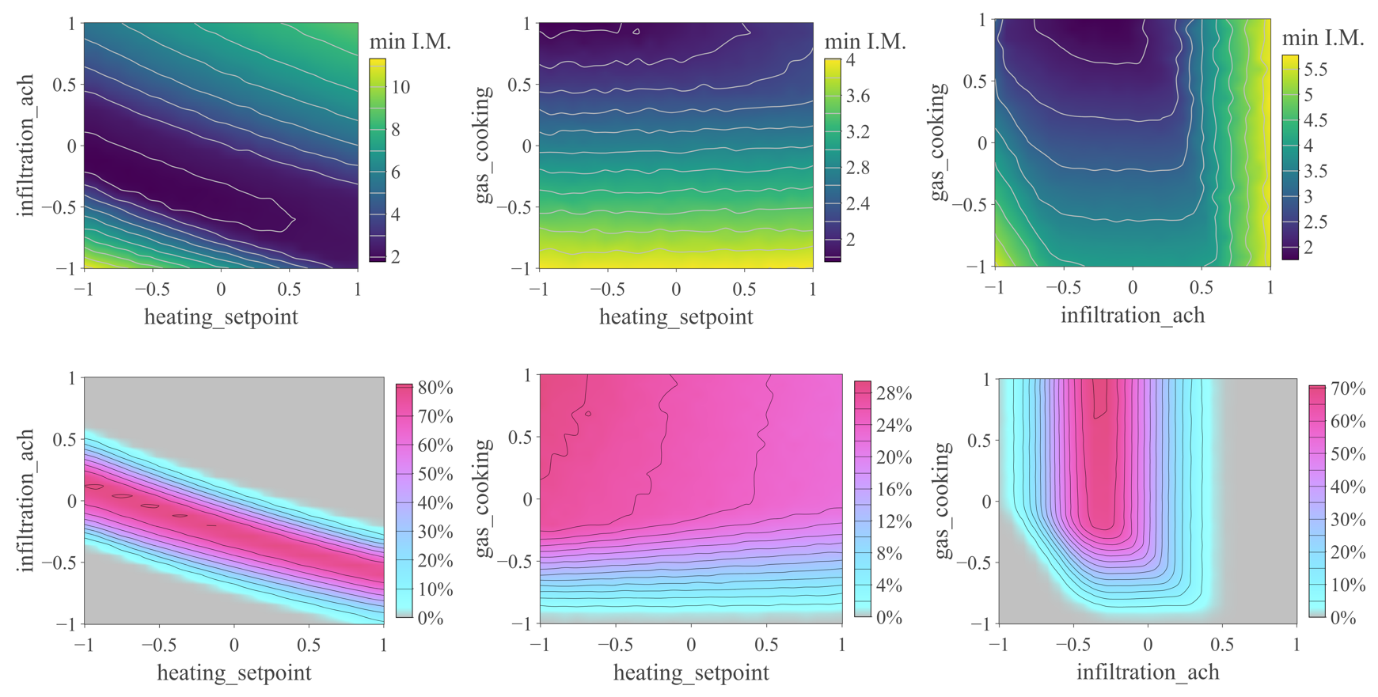
Evaluation of the implausibility measure associated with the new emulator should now allow to bound the non-implausible region more clearly. 

Figure . 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

In accordance with the methodological aim of this work, and to limit its length, we will not discuss results of the above procedure to our example study. However in Section 5.5, albeit driven by a different motivation, we discuss a case where additional runs are selected based on the results of a first emulator, and these are used to train a new emulator and compute the associated IMs.

### Sensitivity to Model Discrepancy magnitude

The plots in Figure 4 concern the case where the magnitude of model discrepancy is 10%. Figure 5 shows MI and OD plots for the same three variables, in the case of 20% MD. Roughly, similar patterns to the ones of Figure 4 emerge, albeit spread over larger regions: due to a lower confidence in the model, it is now more difficult to rule out points as implausible. The percentage of non-implausible points has risen to a remarkable 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 3 shows, we have remarkably precise emulators. Hence, almost 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 3.1.

### Role of different months

The trained emulators allow to explore and answer a wide range of questions, for which it would be otherwise impossible to draw sound inference. We briefly discuss one such example here, by looking into the compatibility between the imposed constraints (the measured consumptions) and the correlation of model predictions across different months. Note, this is in fact an important question for calibration: a strong correlation between the modelled consumption of two months will induce constraints on the pairs of observed consumptions for which a match is possible.

Ideas on what to report. Let me know how that sounds.

* Have a 9x9 scatter plot showing the correlation of the modelled consumptions between all pairs of months. To each plot, can overlap the one point identifying observed consumptions for the pair of months in question.
* Report, for each month, the contribution it gives to cut out space, once the other 8 months are accounted for (eg, Jan/Feb almost interchangeable, Sept cuts a lot).

## Do it in more waves: 100 + 100 > 1000?

In this section:

* Pretend to only have 100 points, and use them for training+validation. Then select other 100 of the original runs, in a neighbourhood of the non-implausible region. How that reduces uncertainty? Does it even reduce it wrt the original case where 1000 runs were used?

# Discussion and Conclusions

References

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

2. Royapoor, M., A. Antony, and T. Roskilly, *A review of building climate and plant controls, and a survey of industry perspectives.* Energy and Buildings, 2017.

3. Raftery, P., M. Keane, and J. O’Donnell, *Calibrating whole building energy models: An evidence-based methodology.* Energy and Buildings, 2011. **43**(9): p. 2356-2364.

4. Daly, D., P. Cooper, and Z. Ma, *Understanding the risks and uncertainties introduced by common assumptions in energy simulations for Australian commercial buildings.* Energy and Buildings, 2014. **75**(Supplement C): p. 382-393.

5. International Organization for Standardization, *ISO 6946:2007: Building components and building elements -- Thermal resistance and thermal transmittance -- Calculation method.* 2007.

6. CIBSE, *Guide A; Environmental Design*. 2015, The Chartered Institution of Building Services Engineers: London.

7. American Society of Heating, R.a.A.-C.E., *2017 ASHRAE Handbook—Fundamentals*. 2017, Atlanta, GA.

8. International Organization for Standardization, *ISO 9869-1:2014: Thermal insulation Building elements (In-situ measurement of thermal resistance and thermal transmittance).* 2014.

9. Rasooli, A., L. Itard, and 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, 2016. **119**(Supplement C): p. 51-61.

10. Deconinck, A.-H. and S. Roels, *Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements.* Energy and Buildings, 2016. **130**(Supplement C): p. 309-320.

11. Meng, X., et al., *Factors affecting the in situ measurement accuracy of the wall heat transfer coefficient using the heat flow meter method.* Energy and Buildings, 2015. **86**(Supplement C): p. 754-765.

12. Ficco, G., et al., *U-value in situ measurement for energy diagnosis of existing buildings.* Energy and Buildings, 2015. **104**(Supplement C): p. 108-121.

13. Gaspar, K., M. Casals, and M. Gangolells, *A comparison of standardized calculation methods for in situ measurements of façades U-value.* Energy and Buildings, 2016. **130**(Supplement C): p. 592-599.

14. Desogus, G., S. Mura, and R. Ricciu, *Comparing different approaches to in situ measurement of building components thermal resistance.* Energy and Buildings, 2011. **43**(10): p. 2613-2620.

15. Hoffmann, C. and A. Geissler, *The prebound-effect in detail: real indoor temperatures in basements and measured versus calculated U-values.* Energy Procedia, 2017. **122**(Supplement C): p. 32-37.

16. Marshall, A., et al., *Domestic building fabric performance: Closing the gap between the in situ measured and modelled performance.* Energy and Buildings, 2017. **150**(Supplement C): p. 307-317.

17. P. Baker, *U-values and Traditional Buildings, Historic Scotland ConservationGroup,* . 2011: Glasgow.

18. Royapoor, M. and T. Roskilly, *Building model calibration using energy and environmental data.* Energy and Buildings, 2015. **94**: p. 109-120.

19. Yassaghi, H., N. Mostafavi, and S. Hoque, *Evaluation of current and future hourly weather data intended for building designs: A Philadelphia case study.* Energy and Buildings, 2019. **199**: p. 491-511.

20. Hong, T., et al., *Advances in research and applications of energy-related occupant behavior in buildings.* Energy and Buildings, 2016. **116**: p. 694-702.

21. Tian, W., et al., *A review of uncertainty analysis in building energy assessment.* Renewable and Sustainable Energy Reviews, 2018. **93**: p. 285-301.

22. Macdonald, I.A., *Quantifying the effects of uncertainty in building simulation*. 2002, University of Strathclyde Glasgow.

23. The Office of Gas and Electricity Markets. *Typical Domestic Consumption Values*. 2017 [cited 2017 17 November]; Available from: <https://www.ofgem.gov.uk/gas/retail-market/monitoring-data-and-statistics/typical-domestic-consumption-values>.

24. Energy Saving Trust, *At Home with Water*. July 2013, Energy Saving Trust: London.

25. Energy Saving Trust, *Measurement of Domestic Hot Water Consumption in Dwellings*. 2008, Department for Environment, Food and Rural Affairs,

: London.

26. Kragh, J., et al., *Possible explanations for the gap between calculated and measured energy consumption of new houses.* Energy Procedia, 2017. **132**(Supplement C): p. 69-74.

27. Newcastle University Urban Observatory. *King's Gate weather station*. Available from: <http://www.urbanobservatory.ac.uk/>.

28. ASHRAE, *ASHRAE Guideline 14-2002: Measurement of Energy and Demand Savings (2002)*. 2002.

29. US Department of Energy. *Air Flow Network Multizone Surface* 2018 [cited 2018 31 Jan]; Available from: <https://bigladdersoftware.com/epx/docs/8-5/input-output-reference/group-airflow-network.html#airflownetworkmultizonesurface>

30. Jason Palmer; Ian Cooper, *United Kingdom Housing Energy Fact File*. 2013, Department of Energy and Climate Change: United Kingdom (London).

31. Eguaras-Martínez, M., M. Vidaurre-Arbizu, and C. Martín-Gómez, *Simulation and evaluation of Building Information Modeling in a real pilot site.* Applied Energy, 2014. **114**: p. 475-484.

32. Wei, S., R. Jones, and P. de Wilde, *Driving factors for occupant-controlled space heating in residential buildings.* Energy and Buildings, 2014. **70**: p. 36-44.

33. Rodríguez, G.C., et al., *Uncertainties and sensitivity analysis in building energy simulation using macroparameters.* Energy and Buildings, 2013. **67**: p. 79-87.

34. Guerra Santin, O., L. Itard, and H. Visscher, *The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock.* Energy and Buildings, 2009. **41**(11): p. 1223-1232.

35. Huang, P., G. Huang, and Y. Wang, *HVAC system design under peak load prediction uncertainty using multiple-criterion decision making technique.* Energy and Buildings, 2015. **91**: p. 26-36.

36. Corrado, V. and H.E. Mechri, *Uncertainty and sensitivity analysis for building energy rating.* Journal of Building Physics, 2009. **33**(2): p. 125-156.

37. Wang, L., P. Mathew, and X. Pang, *Uncertainties in energy consumption introduced by building operations and weather for a medium-size office building.* Energy and Buildings, 2012. **53**: p. 152-158.

38. de Wilde, P., W. Tian, and G. Augenbroe, *Longitudinal prediction of the operational energy use of buildings.* Building and Environment, 2011. **46**(8): p. 1670-1680.

39. Baldi, S., et al., *Real-time monitoring energy efficiency and performance degradation of condensing boilers.* Energy Conversion and Management, 2017. **136**: p. 329-339.

40. O’Brien, *Best Practice Guide, Energy Efficiency: Steam, Hot Water and Process Heating Systems* (2015), Sustainability Victoria.

41. Brian Anderson, *Conventions for U-value calculations*. 2006, Building Research Establishment Scotland: BRE Press, Garston, Watford.

42. Hou D., I.G. Hassan, and L. Wang, *Review on building energy model calibration by Bayesian inference*, Renewable and Sustainable Energy Reviews, 2021. **143,** 110930. https://doi.org/10.1016/j.rser.2021.110930.

43. Goldstein M., Huntley N. (2017) Bayes Linear Emulation, History Matching, and Forecasting for Complex Computer Simulators. In: Ghanem R., Higdon D., Owhadi H. (eds) Handbook of Uncertainty Quantification. Springer, Cham. https://doi.org/10.1007/978-3-319-12385-1\_14

44. Bower, R. G., Goldstein, M., & Vernon, I. (2010). Galaxy formation: a Bayesian uncertainty analysis. *Bayesian analysis*, *5*(4), 619-669.

45. Vernon, I., Liu, J., Goldstein, M. *et al.* Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC Syst Biol* **12,** 1 (2018). https://doi.org/10.1186/s12918-017-0484-3

46. Craig P.S., Goldstein M., Seheult A.H., Smith J.A. (1997) *Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments*. In: Gatsonis C., Hodges J.S., Kass R.E., McCulloch R., Rossi P., Singpurwalla N.D. (eds) Case Studies in Bayesian Statistics. Lecture Notes in Statistics, vol 121. Springer, New York, NY. <https://doi.org/10.1007/978-1-4612-2290-3_2>

47. Domingo, Dario, et al. *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): e2019JF005237.

48. Jason L. Loeppky, Jerome Sacks & William J. Welch (2009). *Choosing the Sample Size of a Computer Experiment: A Practical Guide*, Technometrics, 51:4, 366-376, doi:[10.1198/TECH.2009.08040](https://doi.org/10.1198/TECH.2009.08040)

49. McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). *A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code*. *Technometrics*, *21*(2), 239–245. https://doi.org/10.2307/1268522

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

51. Christopher A. Pope, John Paul Gosling, Stuart Barber, Jill S. Johnson, Takanobu Yamaguchi, Graham Feingold & Paul G. Blackwell (2021). *Gaussian Process Modeling of Heterogeneity and Discontinuities Using Voronoi Tessellations*, Technometrics, 63:1, 53-63, doi:[10.1080/00401706.2019.1692696](https://doi.org/10.1080/00401706.2019.1692696)

52. Pukelsheim, Friedrich (1994). *The Three Sigma Rule.* The American Statistician, 48(2), 88-91. <https://doi.org/10.2307/2684253>.

54. Kennedy, M. C., & O'Hagan, A. (2001). *Bayesian calibration of computer models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), 425-464.

55. O’Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, *91*(10-11), 1290-1300.

56. Craig, P. S., Goldstein, M., Rougier, J. C., & Seheult, A. H. (2001). Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, *96*(454), 717-729

57. Edwards, T. L., Nowicki, S., Marzeion, B., et al. (2021). *Projected land ice contributions to twenty-first-century sea level rise.* Nature, 593(7857), 74-82.

58. Coveney, S. and Clayton, R. H. (2018) Fitting two human atrial cell models to experimental data using Bayesian history matching. Progress in Biophysics and Molecular Biology, 139, 43-58.

59. Andrianakis, Ioannis, et al. "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.

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

61. Goldstein, M. and Wooff, D. (2006). Bayes Linear Statistics: Theory and Methods. Wiley.

Table 2: Sequence of uncertainty analysis on building energy models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Field | Model inputs | Input range | Distribution of uncertainty | Reported variation in model output |
| Building energy model | Occupant behaviour | Presence, density, heat gain | Binary and non-binary | No reports found | 30% [31]  4-26% [32] |
| Building envelope thermal properties | U-values, thickness, surface and moisture properties | Range using Mean and SD [33] | Even [22]  Normal [33] | 42%[34] |
| Weather conditions | Wind speed, direction and pressure coefficients, solar irradiance, air humidity and temperatures | Cold, Med, Hot [33] | Normal [35]  Bivariate Normal [36]  Discrete Distribution [33] | -4% to 6.1% [37] |
| Site micro-environment | Wind-pressure coefficient, ground albedo | Range using Mean and SD [33] | Normal [33] | Not reported |
| HVAC | Values assumed for CoP, SEER and η | Best practice, typical | Normal Distribution [22]  Gamma Distribution [38] | -15.3% -70.3% [37] |
| Internal Gains |  | Low, Med, High [33] | Uniform discrete [33] |  |
| Operational regime | Controls and Scheduling of all HVAC, lighting and plug-in items | Good, average and poor practice [37] | Uniform Discrete [37] | -28.7%-79.2% [37] |
| Observational data | Gas [1] | - | - | Normal [22] | n/a |
| Electricity [1] | - | - | Normal [22] | n/a |
| Temp (Kitchen) | - | - | Normal [22] | n/a |
| Temp (master) | - | - | Normal [22] | n/a |
| Notes  [1] The gas and electricity meters’ accuracy were expected to comply with SI 684 (1983) and IEC 62053 respectively that allow +2.5% or −3.5% of compound instantaneous deviations. | | | | | |

Table 1 Parameter inputs for energy model development of the case-study building

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | Uncertainty range |
| Heating | Natural gas boiler serving a radiator central heating system |  |
| Heating setpoint (setback) | 19°C (16°C) | 17.5°C-20.5°C |
| Heating schedule | 02:00-11:00 + 16:00-24:00 |  |
| Ventilation | Natural ventilation (mechanical extract to family bathroom and en suite) |  |
| Ventilation rate | Highly stochastic, controlled by occupants via openable windows |  |
| Gas boiler seasonal efficiency | 65% (15 years old non-condensing gas-fired system boiler – 77°C/55°C F+R) | 60% - 75% |
| DHW consumption | 0.59 litre/m2/day |  |
| Cooling setpoint (setback) | Uncontrolled |  |
| Nominal lighting power density | 1.4 W/m2 (manually controlled) to achieve 200 lux |  |
| Occupants | 2 people in total |  |
| Internal gains[a] | 6 W/m2 |  |
| Gross (conditioned) area | 168.66m2 (148.93m2) |  |
| Observed annual gas (electricity) consumption (2016) | 15,381 kWh (2,991 kWh) |  |
| **Fabric properties:** |  |  |
| Glazing (with low emissivity coating) | 1.788 W/m2K (3mm self-cleaning pane, 20mm Argon filled cavity, 3mm low emissivity pane) | |
| Glazing G Value (solar transmittance) | 0.691 |  |
| External walls [b] ( W/m2K) | 0.544 | ± 15% |
| Roof [c] (W/m2K) | 0.213 | ± 15% |
| Floor [d] ( W/m2K) | 0.335 | ± 5% |
| Infiltration (ac/h) [e] | 0.5 | 0.25 - 0.95 |
| [a] Electricity (ICT and appliances): 3 W/m2; Gas (catering): 3.3 W/m2 | | |
| [b] 100mm brickwork, 50mm Stone wool insulation, 100mm blockwork, 10mm plasterboards | | |
| [c] 25mm Clay tile roofing, loft space, 180mm glass fibre quilt insulation, 10mm plasterboards | | |
| [d] 100mm cast concrete, 7mm screed, 4mm high gauge polythene DPM, 5 mm foil-backed underlay, 15mm solid wood flooring | | |
| [e] Empirical values derived from table 4.16 (CIBSE Guide A) for a two-storey property on normally exposed site | | |

Table 3 Input parameter variations for Bayesian emulator development

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 |
| Description | Heating setpoint [17.5°C-20.5°C] | Boiler seasonal efficiency | External wall U-value | Roof U-Value | Floor U-value | Infiltration rate (ach) | DHW consumption  (L/day/person) | Cooking |
| Base model input (1st wave) | 19°C | 65% | 0.544 | 0.213 | 0.337 | 0.5 | 300 (1st wave)  120 (2nd wave) | 3% of total domestic energy use |
| Base model input (2nd wave) | 17.5°C | 65% | 0.544 | 0.213 | 0.337 | 0.5 | 120 |  |
| Range of variation | ± 1.5°C | 60% – 75% | ± 15% | ± 15% | ± 5% | 0.25 to 0.95 | 70-250 L/day | 1.05% - 6.3% |
| Rational | [a] | [b] | [c] | [c] | [d] | [e] | [f] | [g] |
| Uncertainty quantification | Forward | Forward | Forward | Forward | Forward | Forward | Inverse | Inverse |
| Element varied in E+ batch simulations | Heating setpoint | Boiler seasonal efficiency | Wall cavity Insulation thickness [ 40mm-63mm] | Insulation thickness [150mm -210mm] | Insulation thickness [45mm- 55mm] | [e] |  |  |
| [a] Manufacturer’s room thermostat resolution reported at ± 0.5°C with an additional ± 1°C allowed for time-dependent drift degradation.  [b] Boiler insulation, heat exchanger and working fluid degradation, limescale and total dissolved solids leading to an accumulated min and Max performance degradation of 4% to 23% [39, 40]. These levels of degradation were imposed on boiler manufacturer’s quoted efficiency of 78%  [c] Although most literature report in-situ wall and roof measurements to be better than elemental method calculation suggestions [14, 16, 17], an equally distributed ± 15% imposed to first cater for all eventualities and enable the uncertainty emulator to assess the entire Latin hypercube space (including worst scenario range).  [d] as per [c] although the magnitude of variations reported for floors were smaller than those of walls/roofs [17] and non-suspended ground floors with no air cavities have much greater thermal unity [41] so a tighter band of ± 5% was imposed to reflect literature findings.  [e] As outlined in the last paragraph of sections 2 and 4.  [f] From field measurements of DHW consumption in the UK [25] where the mean DHW consumption per person in the UK is reported as 122 litres/day ± 18 litres/day (i.e. ±15% variation) leading to mean DHW energy consumption of 16.8 MJ/day ± 2.2 MJ/day (95% statistical confidence). In the 1st wave the model input was a much larger values of >300l/day and 53MJ/day, However the model predictions were calibrated to return close results to the observed energy consumption for the case-study building given its high fossil fuel consumptions.  [g] Cooking has been observed to currently account for an average of 3% of total household energy demand with historical data also indicating a maximum of 6% [30]. This observed data informs the average and maximum cooking demand with 1% also selected by the authors to represent a probable lower boundary. | | | | | | | | |

1. In R, this is achieved by poly(X, deg=2), where X is the matrix containing a training point in each row. For 8 variables, there is a total of 44 linear, quadratic and interaction terms. [↑](#footnote-ref-1)