First-Wave Emulators

Contents

First-Wave Emulators

This document relates to the project with Mohammad Royapoor on the modeling of a building's energy consumption. I first recall what the setting and aim are, then detail the choices I have made to construct a first wave of emulators, and finally explore the history matching under some scenarios.

1 The Simulator: Inputs and Outputs

The available simulator is called Energy+. It is used to predict the energy consumption of a building whose shape and dimensions are passed to the model. The building we have simulation results for is Mohammad's house. The simulator can be though of as taking in input a number of parameters (*e.g.*, air permeability of walls, efficiency of energy devices, etc) and a year-long sequence of daily routine profiles (*e.g.*, when during the day, and for how long, hobs/hot water/etc are used), and accordingly predicting the monthly energy consumption of the building, for all 12 months.[1]

The results of a set of $n = 1000$ runs are available: I may refer to these as to the design runs. The sequence of routing profiles is the same for all n of them. Eight parameters are instead varied from one run to the other: for simplicity, I am going to name these $V_1, \ldots, V_8$. Their physical meaning and the range that each of them covers across the n simulations are shown in Table 1. The n design points in $\mathbb{R}^8$ constituting the experimental design were selected by Hailiang via Latin hypercube sampling, in the 8-dimensional "cube" with side ranges as in Table 1.

For each of the 12 month, we therefore have the n simulated consumptions at the design points; in addition, we have the observed piece of data for the monthly consumption.

[1] I will only consider Gas. While results concerning electricity are also available, they show little to no variability among the design runs.

Table 1: Meaning and range of the eight parameters varied among the n simulations of this work. *(Mohammad may provide more appropriate descriptions for the middle column.)*

Short Name	**Physical Meaning**	**Range**
V_1	Heating Setpoint	$[17.5, 20.5]$ °C
V_2	Boiler Efficiency	$[0.6, 0.75]$
V_3	External wall thickness	$[4, 6.3]$ cm
V_4	Roof quilt thickness	$[15, 21]$ cm
V_5	Floor insulation thickness	$[4.5, 5.5]$ cm
V_6	Infiltration rate	$[0.2, 0.95]$ ac/h
V_7	DHW Consumption	$[6.15, 22] \times 10^{-6}$ litre/day
V_8	Cooking	$[1.05, 6.3]$ W/m^2

Figure 1 shows, for each month, the distribution of simulated consumptions and the location of the observed consumption. For the three summer months (June, July, August), the observed consumption is (often far) greater than any of the n simulated consumptions. In the following, I am going to exclude these months from the analysis – we can discuss and come up with better ideas if needed. (*Note: the rest of the document is mostly written in first person plural, as in a slightly more paper-ish style.*)

2 Construction of the Emulators

For each of the nine months we consider, it is natural to see the simulator as a function

$$\begin{array}{rccc} f\colon & \mathbb{R}^8 & \longrightarrow & \mathbb{R}^+ \\ & \boldsymbol{x} & \mapsto & y \end{array}.$$

To any choice $\boldsymbol{x} \in \mathbb{R}^8$ of the eight variables reported in Table 1, the function associates the simulated gas consumption $y = f(\boldsymbol{x})$ for that month. In the following, we first set the notation and introduce the parameters which will need to be estimated in order to build an emulator of such a function (Section 2.1). The procedure used to estimate/choose these parameters for each of the nine months of interest is detailed in Section 2.2.

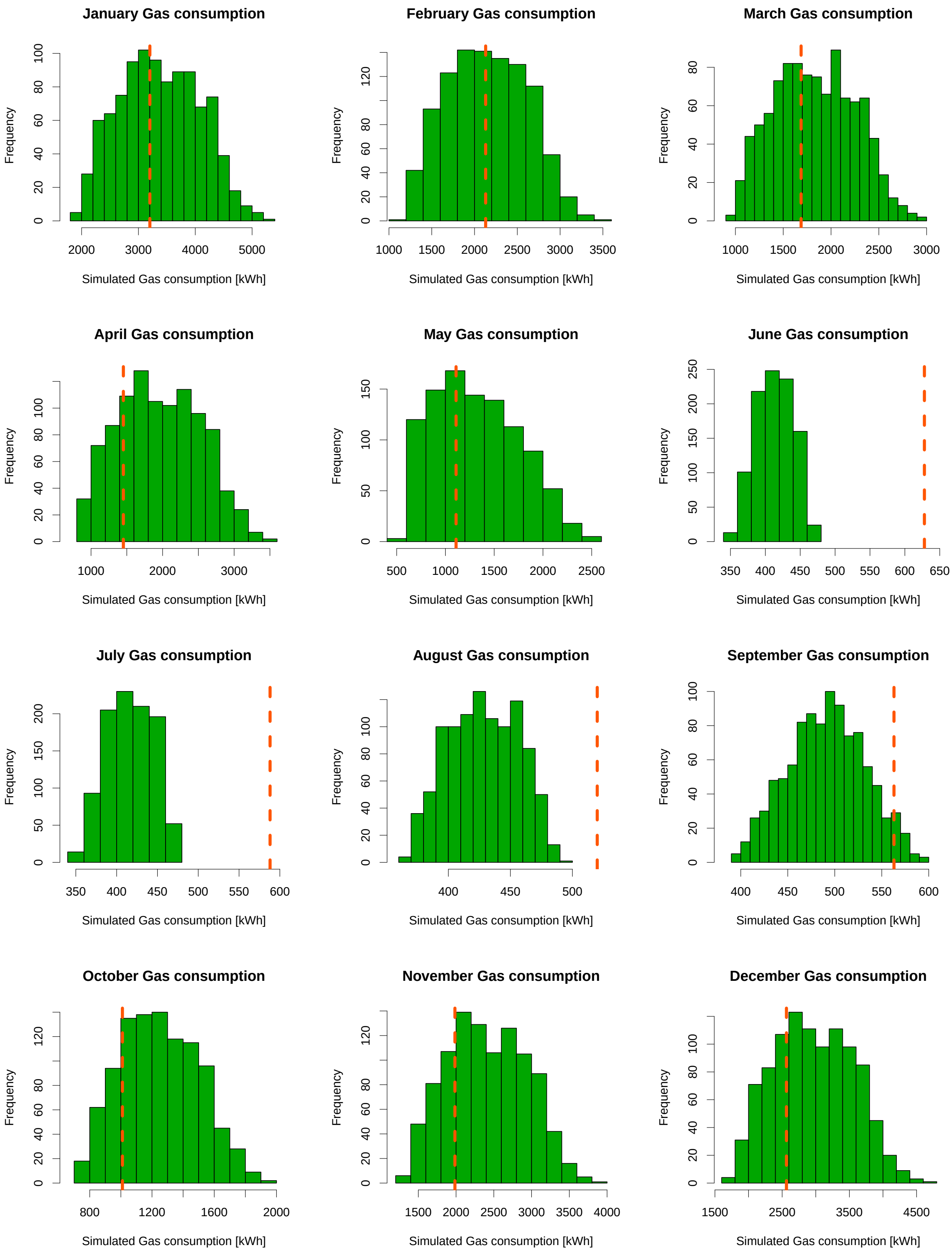

Figure 1: Distribution of the simulated monthly gas consumptions corresponding to the experimental design used in this work ($n = 1000$ simulations). The x-value of the vertical dashed line in each plot locates the observed consumption for that month.

2.1 General Setting

The $n = 1000$ runs provide the value of the function $f(\cdot)$ at the design points, which we denote as $\boldsymbol{x_i} \in \mathbb{R}^8$, $i = 1, \dots, n$. Following the classical emulation approach, we model $f(\cdot)$ as a stochastic process, specifically in the following way:

$$f(\boldsymbol{x}) \;=\; \sum_{j=0}^{r} \beta_j \, g_j(\boldsymbol{x}^{[A]}) \;+\; \eta(\boldsymbol{x}^{[A]}) \;+\; \varepsilon(\boldsymbol{x})\,, \qquad (1)$$

where the first term (the sum) represents the mean of $f(\cdot)$, while the other two terms are both zero-mean stochastic processes who account for the local deviation of $f(\cdot)$ from its mean. More details follow.

- The vector $\boldsymbol{x}^{[A]} \in \mathbb{R}^p$, $p \leq 8$, is meant to denote the "active" inputs of the function $f(\cdot)$, the ones which have a more notable effect in explaining the variability of $f(\cdot)$ across different runs. The mean of $f(\boldsymbol{x})$ is a linear combination of functions of the active inputs $\boldsymbol{x}^{[A]}$: these functions are the $g_j(\cdot)$, which we will refer to as regressors. The function $g_0(\cdot)$ will be chosen to be identically equal to 1, so the mean of $f(\boldsymbol{x})$ is an affine combination of the regressors $g_1(\boldsymbol{x}^{[A]}), \dots, g_r(\boldsymbol{x}^{[A]})$.

- The process $\eta(\cdot)$ is a zero-mean stochastic process, with squared-exponential covariance function. That is (dropping for a moment the superscript $[A]$ for convenience), we have:

  $$\mathrm{Cov}\big[\eta(\boldsymbol{x}), \eta(\boldsymbol{x}')\big] = {\sigma_\eta}^2 \, \exp\left(-\frac{1}{2} \sum_{k=1}^{p} \left(\frac{x_k - x'_k}{d_k} \right)^2 \right), \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p. \qquad (2)$$

  This form assumes that the variance of $\eta(\cdot)$ is constant across the space, equal to ${\sigma_\eta}^2$. The values of both ${\sigma_\eta}^2$ and the correlation lengths d_k will be set according to the case (Section 2.2).

- The term $\varepsilon(\boldsymbol{x})$ represents a "nugget" term, accounting for the residual local variability of $f(\cdot)$ not accounted for by the previous term. As it is usual, we model this as a zero-mean real stochastic process independent of $\eta(\cdot)$, with constant variance ${\sigma_\varepsilon}^2$ and uncorrelated outputs at any two points of the space.

The previous specifications uniquely determine the mean and covariance structure of the process $f(\cdot)$ in (1). In light of the simulation outputs

$$y_i = f(\boldsymbol{x_i}), \quad i = 1, \dots, n,$$

we can use the linear Bayes methodology to adjust the mean and covariance specifications above. (I won't go into the details of presenting the formulas here, we'll do this for the paper instead, according to the journal we'll submit the work to).

2.2 Specific Choices

This section provides the details of the specific choices made to build the nine emulators of this work (each associated with one of the non-summer months). In particular, it concerns the choices of the active inputs $\boldsymbol{x}^{[A]}$, regressors $g_j(\cdot)$, correlation lengths d_k and variances ${\sigma_\eta}^2$, ${\sigma_\varepsilon}^2$.

To select the active variables and the associated regressors, we consider a larger set of potential regressors formed by all linear, quadratic and interaction terms of the eight input variables (after linearly rescaling these onto the range $[-1, 1]$). This yields a total of 44 potential regressors. Hence, for each integer $N_{\text{regr}} \leq 44$, among all linear models expressing y_i as linear combination of exactly N_{regr} regressors, we select the model with the greatest coefficient of determination R^2. This procedure allows us to identify, for each integer N_{regr}, the set of N_{regr} regressors of order at most 2 which best explain the set of simulated gas consumptions $\{y_i\}_{i=1,\ldots,n}$.

The results of this procedure show that, for all nine months, the "best" five regressors are sufficient to explain more than 99% of the output variance; in fact, with the only exception of September, four regressors are already sufficient to achieve the result. In spite of this already excellent fit, adding sequentially further regressors to the models, as per the above procedure, improves the models significantly. For all the nine months, and up to models with 13 regressors, the null hypothesis of a zero coefficient associated with any of the considered regressor is confidently rejected with a p-value lower than the machine epsilon ($\approx 10^{-16}$). In order to avoid a model with an excessive number of regressors, we use the procedure above to select the first 10 best regressors for each month, and use these as functions $g_1(\cdot), \ldots, g_r(\cdot)$ in equation (1), with $r = 10$ and $g_0(\cdot) \equiv 1$. An exception to this rule is represented by the emulator for the month of May: this is discussed further below.

Once the set of regressors for a given month has been chosen, we consider as active inputs the ones in the set $\{V_1, \ldots, V_8\}$ which play an active role in the regressors. In a few cases, one variable V_j not appearing in the expression of the selected regressors may have nonetheless been included in the set of active inputs: this is due to the fact that

Table 2: For each of the nine months considered, in order from left to right: i) the list of regressors used as functions $g_j(\cdot)$ in equation (1); ii) the list of active inputs used to specify the covariance function in equation (2); iii) the adjusted R^2 of the linear regression built with the regressors in column i). In this column, the "$*$" symbol is used as shorthand notation for all linear and interaction terms, *e.g.* $a * b * c = \{a, b, c, ab, ac, bc\}$.

Month	**Regressors**	**Active Inputs**	**Adj.** R^2
Jan	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6	0.9998
Feb	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6	0.9998
Mar	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6, V_7	0.9997
Apr	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6	0.9998
May	$V_1{*}V_2{*}V_6$, V_3, V_8, ${V_6}^2$, $V_1{V_6}^2$, ${V_6}^3$	V_1, V_2, V_3, V_6, V_8	0.9991
Sep	$V_1{*}V_2{*}V_6$, V_3, V_4, V_8, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6, V_8	0.9994
Oct	$V_1{*}V_2{*}V_6$, V_3, V_4, V_8, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6, V_8	0.9996
Nov	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6, V_8	0.9998
Dec	$V_1{*}V_2{*}V_6$, V_3, V_4, ${V_2}^2$, ${V_6}^2$	V_1, V_2, V_3, V_4, V_6	0.9998

that variable does appear as significant in the corresponding linear regression model, as a term just beyond the tenth. It does therefore play an important role in explaining the structure of the regression residuals, through which we fit a stochastic process (the terms $\eta(\cdot) + \varepsilon(\cdot)$ in equation (1)). Table 2 shows the sets of regressors and active inputs selected for each month, alongside the adjusted R^2 of the associated linear regression.

As mentioned previously, a slightly different procedure has been used to select the regressors and active inputs for the month of May. For this month, the plot of the linear regression residuals (linear model built as above with the best 10 regressors of order 2) against V_1 and especially V_6 showed a remarkable patter, resembling the one of a polynomial of order higher than 2. In light of this, before running the above linear regression procedure, we include all cubic terms involving either V_1 or V_6 in the set of potential regressors. Within the first 11 terms, both the cubic term in V_6 and the interaction term between V_1 and ${V_6}^2$ are picked by the procedure. This inclusion makes

Table 3: Month by month, values of the correlation length d $(= d_1 = \ldots = d_p)$ used in the squared-exponential covariance function (2), and of the residuals' standard deviation of the linear model built with the regressors in Table 2.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Sep**	**Oct**	**Nov**	**Dec**
d	0.35	0.3	0.35	0.3	0.4	0.4	0.45	0.5	0.35
σ [kWh]	9.26	7.25	7.19	7.11	12.69	0.96	4.91	7.59	8.26

the relationship between the new model residuals and the two input variables much less obvious to interpret. We therefore consider this set of 11 regressors for May, and choose the variables appearing in them as active inputs. See Table 2 for details.

Once the regressors and active inputs are chosen for all the months, values for the correlation lengths d_k and the two "prior" variances ${\sigma_\eta}^2, {\sigma_\varepsilon}^2$ have to be set in order to build an emulator. As seen by equation (1), the quantity ${\sigma_\eta}^2 + {\sigma_\varepsilon}^2$ accounts for the local variability displayed by $f(\cdot)$ once the mean has been removed. For this reason, we impose:

$$ {\sigma_\eta}^2 + {\sigma_\varepsilon}^2 = \sigma^2 \, , \tag{3}$$

where σ^2 is the residuals' variance of the linear model used to select the regressors in $f(\cdot)$. Specifically, we impose that 95% of the total variability σ^2 is ascribed to the "structured" process $\eta(\cdot)$ and the remaining 5% to the "random noise" process $\varepsilon(\cdot)$:

$$ {\sigma_\eta}^2 = 0.95\,\sigma^2 \, , \qquad {\sigma_\varepsilon}^2 = 0.05\,\sigma^2 \, . \tag{4}$$

As far as the correlation lengths d_k are concerned ($k = 1, \ldots, p$, equation (2)), in this first stage we make the simplifying assumption that, for a given month, they are all equal to each other:

$$ d_1 = \ldots = d_p = d \, . \tag{5}$$

For a fixed number p of terms in equation (2), increasing the value of d leads to an increase in the correlation between different outputs of the process $\eta(\cdot)$. Notice, however, that for a fixed d the correlation decreases if more terms are accounted for in the sum of equation (2). According to the month considered and to the number of active inputs p, we choose a value of d in the range $[0.3, 0.5]$. Table 3 reports the specific choices of d for each month, alongside the value σ^2 used as prior variance for the model. The predictive ability of the emulators built with the parameters detailed in this section is tested via leave-one-out cross-validation, as the next section explains.

2.3 Validation of the Emulators

The emulators built via the choices detailed in Section 2.2 are validated via leave-one-out cross-validation. The method consists in training an emulator, with the choices of the previous section, on all pairs $(\boldsymbol{x_i}, y_i)$ but one, and then using the emulator to predict the output associated with the left-out input point. By leaving out all pairs in turn and comparing the emulator prediction to the known output, we can assess the predictive ability of the emulator.

Suppose the pair $(\boldsymbol{x_i}, y_i)$ is left out from the training set of the emulator. Let $\widehat{y}_i$ and $\widehat{\sigma}_i$ be the emulator mean prediction and standard deviation at the point $\boldsymbol{x_i}$, respectively. The following quantity:

$$\widehat{\varepsilon}_i = \frac{\widehat{y}_i - y_i}{\widehat{\sigma}_i} \tag{6}$$

measures the distance between the emulator prediction and the real simulator output y_i, in multiples of the emulator standard deviation. We therefore call $\widehat{\varepsilon}_i$ the (cross-validated) standardised error for the i^{th} data point. Under Pukelsheim 3σ-rule, a well-calibrated emulator would show at least 95% of the $\widehat{\varepsilon}_i$ to be in modulus smaller than 3.

Figure 2 shows, for each of the months of interest, the distribution of the $n = 1000$ standardised errors. The histograms suggest that all nine emulators are able to provide accurate predictions and correctly assess the uncertainty of these. In addition, in Figure 3 we show the plots of the same set of standardised errors against the emulator fitted values. The plots aim to assess whether the errors' distribution is mostly coherent across the different $\widehat{y}_i$ values. Most of the plots confirms that this is the case. In a few rare cases (*e.g.*, April), the magnitude of the standardised errors may be higher near the edges of the range of simulated values. Overall, the plots show nonetheless a satisfactory fit, especially for values $\widehat{y}_i$ in a neighbourhood of the observed gas consumption for that month: this is represented as a vertical orange dashed line in each panel.

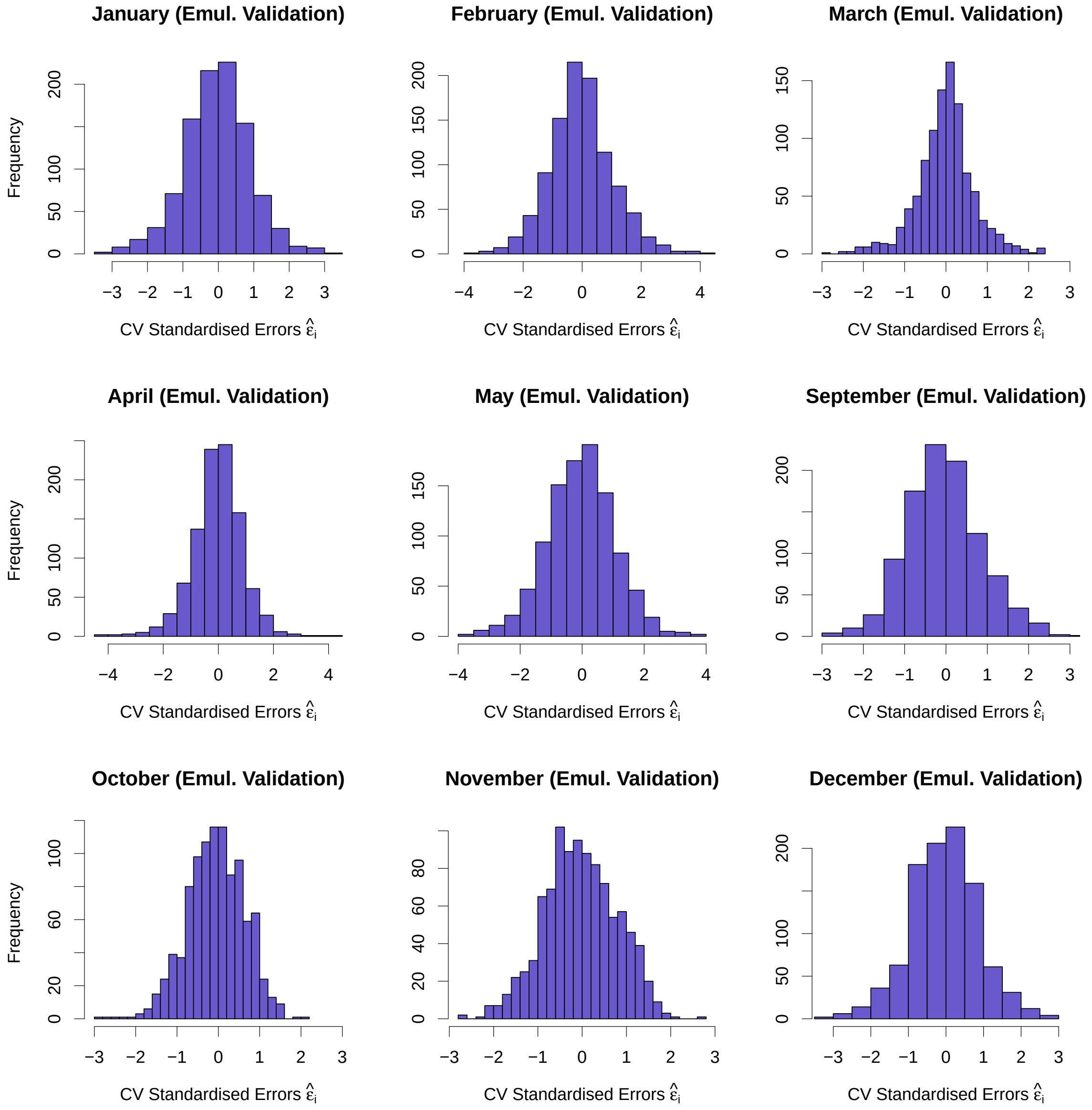

Figure 2: Validation of the emulators. In each plot, the histogram of the $n = 1000$ cross-validated standardised errors $\widehat{\varepsilon}_i$ is shown (equation (6)).

3 History Matching

The emulators built in Section 2 allow to predict, for any input $\boldsymbol{x} \in \mathbb{R}^8$, the simulated gas consumption $f(\boldsymbol{x})$ associated with that set of inputs. Since observed gas consumptions are available for each month, we can use the emulators to select inputs $\boldsymbol{x}$ whose simulated gas consumptions are “compatible” with the observed data. The context and notation

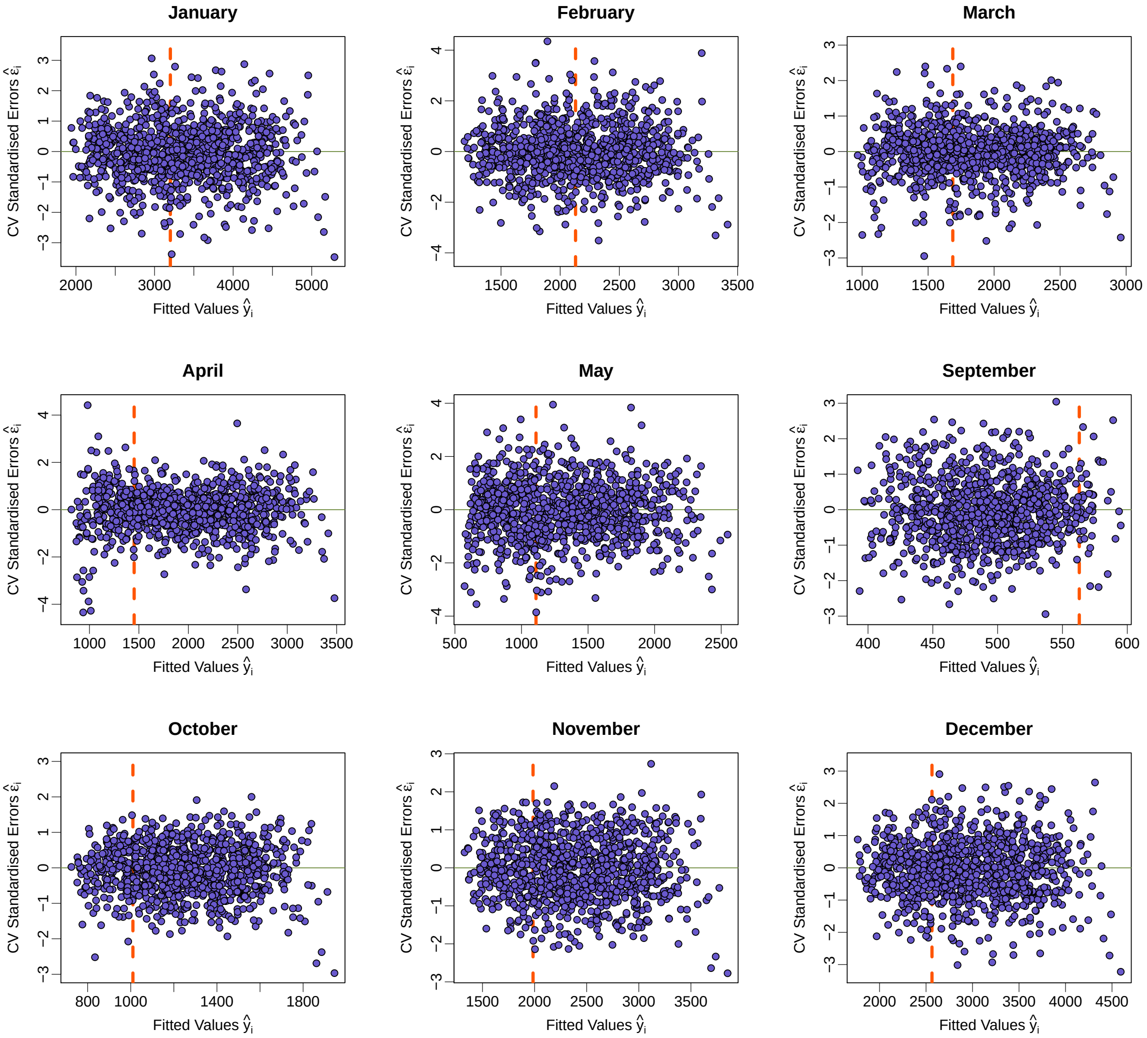

Figure 3: Validation of the emulators. For the different months, the standardised errors in (6) are plotted against the emulator's fitted values. In each plot, the vertical dashed line identifies the x-value at which observed gas consumption for the month is available.

to carry out the comparison are introduced below.

3.1 Notation and Choices

As in the rest of this work, $f(\boldsymbol{x})$ will denote the simulated gas consumption associated with input $\boldsymbol{x}$. With a slight abuse of notation which should however cause no confusion, we denote by $\mathbb{E}[f(\boldsymbol{x})]$ and $\mathrm{Var}[f(\boldsymbol{x})]$ the emulator mean prediction and variance associated with input $\boldsymbol{x}$. Finally, we denote by $z \in \mathbb{R}$ the observed gas consumption. Note that, albeit not explicitly signaled in the notation, all these quantities depend on the particular month considered.

When using the emulator to assess the compatibility between the (unknown) value $f(\boldsymbol{x})$ corresponding to a given input $\boldsymbol{x}$, and the available observed consumption z, we need to account for at least three different types of uncertainties.

1. The emulator prediction is only an approximation to the actual model response $f(\boldsymbol{x})$: this uncertainty is naturally quantified by the emulator standard deviation associated with the prediction.

2. Even if the emulator were a perfect reproduction of the simulator, the latter would only be an imperfect reproduction of reality. We ultimately want to learn about the real world, so model discrepancy (MD) should be accounted for.

3. Due to unavoidable measurement errors (MEs), the observed consumption z differs from the actual amount of gas consumed in the building during the month.

The aim of the first wave of emulators is to discard those inputs $\boldsymbol{x}$ which, once the previous uncertainties are accounted for, are extremely unlikely to generate an actual consumption z in the real world. In order to quantify the mismatch between an input $\boldsymbol{x} \in \mathbb{R}^8$ and the observed z for the month M, we consider the following implausibility measure[2]:

$$I_M(\boldsymbol{x}) = \frac{\left|\, \mathbb{E}[f(\boldsymbol{x})] - z \,\right|}{\sqrt{\mathrm{Var}[f(\boldsymbol{x})] + \mathrm{Var}[\epsilon_{\mathrm{MD}}(\boldsymbol{x})] + \mathrm{Var}[\epsilon_{\mathrm{ME}}(z)]}}\,, \qquad \boldsymbol{x} \in \mathbb{R}^8. \tag{7}$$

The subscript M denotes the dependence on the month considered (it is omitted from the terms on the right-hand side for simplicity of notation). The quantity $I(\boldsymbol{x})$ is dimensionless. The last two terms of the denominator represent the uncertainty ascribed to model discrepancy and measurement errors respectively, and have the same dimension as $\mathrm{Var}[f(\boldsymbol{x})]$.

In our case, we assess the measurement error to be of the order of 5% of the reported

[2] I don't add much here, but clearly in a paper we would reference Michael's work and possibly restructure the exposition.

gas consumption z[3]. To translate this into a value for $\mathrm{Var}[\epsilon_{\mathrm{ME}}]$, we make the simplifying assumption that the actual consumption has uniform probability of being anywhere in the interval $[0.975z, 1.025z]$. Since the variance of a uniform random variable on an interval of length L is $L^2/12$, the assumption yields:

$$\mathrm{Var}[\epsilon_{\mathrm{ME}}(z)] = \frac{z^2}{4800}\,. \tag{8}$$

As far as model discrepancy is concerned, we consider three different scenarios, where the discrepancy between a simulated output y and reality is estimated in about 5%, 10% and 20% of y respectively. Since the output y is not available for a general input $\boldsymbol{x}$, we consider the quantity $\mathbb{E}[f(\boldsymbol{x})]$ as its surrogate, therefore computing $\mathrm{Var}[\epsilon_{\mathrm{MD}}(\boldsymbol{x})]$ as

$$\frac{\mathbb{E}[f(\boldsymbol{x})]^2}{4800}\,, \quad \frac{\mathbb{E}[f(\boldsymbol{x})]^2}{1200}\,, \quad \frac{\mathbb{E}[f(\boldsymbol{x})]^2}{300} \tag{9}$$

in the three scenarios respectively.

For the month M, we define an input $\boldsymbol{x}$ as "non-implausible for M" if the value $I_M(\boldsymbol{x})$ is lower than a given threshold T. Hence, it is natural to define an input as "non-implausible" if the previous condition holds true for all the months. That is, if

$$\max_{M \in \mathcal{M}} \left\{ I_M(\boldsymbol{x}) \right\} < T\,, \tag{10}$$

where $\mathcal{M}$ is the set of nine months considered in this work. Since the previous equation represents the simultaneous imposition of nine conditions, we consider two relatively conservative values of T in equation (10): $T = 4$ and $T = 5$.

3.2 Results

In order to explore the compatibility between different regions of the space and the observed data, we generate a quasi-random sample of $N = 6 \times 10^7$ points in the 8-dimensional cube $C = [-1, 1]^8$ (all inputs have been rescaled onto the range $[-1, 1]$ prior to the analysis, see Section 2.2). The sample is generated via a Sobol sequence.

We start by examining the results when $T = 4$. In the case where only 5% model discrepancy (MD) is considered, none of the 60 million points is classified as non-implausible according to condition (10). Further inspection also reveals that no point can be found

[3] Both for this and for the model discrepancy, Mohammad should soon provide quantitative indications. The results will be adjusted once I get these, before I select the inputs of the set of new runs.

Table 4: Percentages of non-implausible space when accounting for different size of model discrepancy (MD). Results shown for two different values of the threshold T. In the "All" column, condition (10) is used to classify a point as non-implausible. The looser condition $I_{\mathrm{M}}(\boldsymbol{x}) < T$ is instead used in each of the remaining nine columns, for the different months "M" (percentages rounded to the nearest integer). The "≈ 0" value in the table corresponds to 1.63×10^{-4} % of the space.

		All	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Sep**	**Oct**	**Nov**	**Dec**
$T = 4$	5% MD	0	24	24	21	13	14	27	20	19	23
	10% MD	0.33	38	38	33	21	22	45	32	31	36
	20% MD	19.53	70	69	63	39	42	81	59	57	67
$T = 5$	5% MD	≈ 0	30	30	26	16	18	35	25	24	28
	10% MD	3.36	27	48	42	26	28	58	39	38	45
	20% MD	37.03	84	81	78	50	54	92	74	72	82

which matches any eight of the nine months simultaneously. However, by accounting for larger model discrepancy, non-implausible inputs can be found: they constitute 0.33% of the space when 10% MD is accounted for, and 19.5% of the space when MD is increased to 20%.

In Table 4 we report these percentages, together with the percentages of space which are deemed non-implausible when compatibility with a single month is imposed. The percentages corresponding to the case $T = 5$ are also shown. We see that the non-implausible fraction of space corresponding to a 5% MD is essentially zero also in the $T = 5$ case. If we accept that observed consumptions come with a measurement error of at most 5%, then this suggests that the discrepancy between the simulator and reality is likely to be higher than 5%.

As easily foreseeable from the plots of Figure 1, Table 4 reveals that notable parts of the space result non-implausible when compared with the observation of a single month, although the insersection of these parts of the space may be (almost) empty, as in the two 5% MD cases. With the exception of these two rows, however, we can notice the following: the fraction of non-implausible space is about 2 orders of magnitude higher than the product of non-implausible fractions with respect to the single months. This suggests indeed that the nine conditions imposed in equation (10) are not independent of each other, but rather positively correlated.

Note for Michael and Hailiang:

I could analyse the results of the simulations only starting from yesterday (Friday), so I had no time to look for more hidden patterns or for the geometrical distribution of non-implausible points within the 8D space. That is obviously something to explore, perhaps once Mohammad gives us hints on the magnitude of model discrepancy and measurement errors.