PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Volodina V, Challenor P. 2021 The importance of uncertainty quantification in model reproducibility. *Phil. Trans. R. Soc. A* **379**: 20200071.
https://doi.org/10.1098/rsta.2020.0071

Accepted: 21 September 2020

One contribution of 15 to a theme issue 'Reliability and reproducibility in computational science: implementing verification, validation and uncertainty quantification *in silico*'.

Subject Areas:
statistics, mathematical modelling, applied mathematics, computer modelling and simulation, software

Keywords:
emulation, Bayesian methods, error estimates

Author for correspondence:
Victoria Volodina
e-mail: vvolodina@turing.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5253498.

THE ROYAL SOCIETY PUBLISHING

The importance of uncertainty quantification in model reproducibility

Victoria Volodina[1] and Peter Challenor[1,2]

[1]The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QE, UK

VV, 0000-0003-4490-8777; PC, 0000-0001-8661-2718

Many computer models possess high-dimensional input spaces and substantial computational time to produce a single model evaluation. Although such models are often 'deterministic', these models suffer from a wide range of uncertainties. We argue that uncertainty quantification is crucial for computer model validation and reproducibility. We present a statistical framework, termed history matching, for performing global parameter search by comparing model output to the observed data. We employ Gaussian process (GP) emulators to produce fast predictions about model behaviour at the arbitrary input parameter settings allowing output uncertainty distributions to be calculated. History matching identifies sets of input parameters that give rise to acceptable matches between observed data and model output given our representation of uncertainties. Modellers could proceed by simulating computer models' outputs of interest at these identified parameter settings and producing a range of predictions. The variability in model results is crucial for inter-model comparison as well as model development. We illustrate the performance of emulation and history matching on a simple one-dimensional toy model and in application to a climate model.

This article is part of the theme issue 'Reliability and reproducibility in computational science: implementing verification, validation and uncertainty quantification *in silico*'.

1. Introduction

A computer model (simulator) is a coded representation of a true process of interest. We treat a computer model as a mathematical function f that takes varying values of input parameters denoted by a vector $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, and returns output $f(\mathbf{x})$. Owing to the vast improvements in the power of computers, these models, in combination with mathematical modelling and past physical data, are used in the analysis of complex physical systems and decision support [1]. For instance, Galform, a model of galaxy formation, is employed to study the behaviour of galaxies in the presence of dark matter [2]. Another example of powerful and complex models are the three-dimensional General circulation models (GCMs) of the atmosphere and ocean, numerical models based on the calculation of the budgets of mass, energy and momentum on a grid of columns on a sphere [3]. These models allow scientists to improve their understanding of the Earth system and to project the future climate state. In particular, simulations and projections produced by the world-leading climate centres' GCMs are analysed and compared in the Coupled Model Intercomparison Project (CMIP) [4]. These model results and comparisons serve as the basis for climate research [5]. In the civil service, analytical models are used in many ways from appraising and evaluating the impact of policy options to planning the current strategy based on future forecasts [6,7]. These examples demonstrate the importance of assessing the model reliability and reproducibility.

Reproducibility may seem an odd concept in computer models. If we take the same deterministic computer code and run it again, we will get the same answer, given some small variation due to computer architecture, different compilers or at least preserve the global features in output behaviour [8]. But if we change one of the model inputs by a small amount, within its error bounds, so it is not a 'real' change, this will change the output. Similarly, if two laboratories use different models of the same process, the results tend to vary. These two phenomena raise a number of important questions about the interpretation and communication of model results. Such issues have featured heavily in press coverage of computer modelling of COVID-19 (see for example [9]).

In this paper, we argue that instead of reporting a single model estimate, i.e. result from a single simulation or a mean of ensemble of runs, the variability in model results should be acknowledged and reported. This variability, or error about a model estimate, represents how confident modellers are in the results produced by their model. To obtain these estimates, modellers need to identify and quantify different sources of uncertainty.

Uncertainty quantification in computer models is important for a number of reasons. Firstly, the analysis of physical processes based on computer models is riddled with uncertainty, which has to be addressed to perform 'trustworthy' model-based inference such as forecasting (predictions) [1]. Secondly, reporting and communicating uncertainties encountered in models is crucial in maintaining credibility in the model and modelling group (a team of climate scientists and modellers). For example, in climate science, models are constantly being improved and developed, therefore maintaining credibility in the future as model-based information improves is critical [10]. In cases where a model is used for decision support in civil services, the reputation of a single state department or even the whole government, depending on the scale of decision, could be on the line. In particular, commissioners of analysis are warned of the potential damage to credibility caused by overconfidence in their analysis [6].

In this paper, we adopt the subjective Bayesian approach to deal with uncertainties encountered in computer models. Goldstein [1] praises the ability of the subjective approach to translate complex uncertainty judgements provided by modellers into a mathematical formulation. The probabilities and conditional probabilities are employed to represent modellers' uncertainty about the quantities of interest and observations about these quantities, respectively. For complex applications such as climate modelling, subjective Bayesian approach is the only way to perform model-based inference about the physical process of interest by combining limited data with expert knowledge. In cases where we have a very large experiment, objective Bayes characterized by 'neutral' knowledge prior distribution could be adopted. However, [1] still

considers such analysis as the approximation to the full subjectivist analysis. Bayesian emulation and history matching are well-established techniques in the uncertainty analysis of computer models. In this paper, we are interested to showcase these approaches to perform model-based inference about the physical process of interest in the presence of uncertainty. In §2, we consider a general classification of different types of uncertainty presented by [11]. In §3a, we introduce the emulator, which is a statistical representation of a computer model behaviour [12–14]. In other words, a fast approximation to the full model, but one which includes an estimate of its own uncertainty. We proceed to combine emulators with statements of uncertainty about the discrepancy between the model and the physical process and about the measurement uncertainty associated with the observation to perform history matching in §3b. History matching is a global parameter search approach used to learn unknown input parameter values [2,15,16]. A model with fitted values could be used to predict the future behaviour of the system [11]. Section 4 demonstrates how these presented methods work for a climate model example. We finish off §5 with a general conclusion and discussion of the importance of these methods for model reproducibility.

2. Sources of uncertainty

In general, uncertainties can be categorized as either aleatory or epistemic [17]. *Aleatoric uncertainty* is associated with the internal variability in the system, and therefore cannot be reduced by conducting more experiments or collecting more data. It is natural to employ a probabilistic framework to model this type of uncertainty. The second type, *epistemic uncertainty*, arises from the deficiency in the model, caused by limited knowledge about the phenomena. Contrary to aleatoric uncertainty, there is a possibility to reduce this type of uncertainty by improving our knowledge about the system.

The process of identifying and classifying various sources of uncertainty is complex and model specific. For instance, in climate science, GCMs suffer from *initial condition uncertainty*, since the climate state generated by a model is sensitive to the changes in the initial state [10]. To analyse and represent this type of uncertainty, the projections (long-term predictions) from a climate model are computed from the ensemble of possible initial conditions [18]. Another example is in government where analysts and modellers identify a special type of uncertainty, *deep uncertainty*, which corresponds to all the events whose impacts on the policy outcome are not clear [6].

In this paper, we adopt the classification of sources of uncertainty provided by [11] and common to a wider class of models. Figure 1 depicts the main types of uncertainty encountered when dealing with computer model $f(\mathbf{x})$, a representation of physical process y. *Parameter uncertainty* indicates that we have a number of unknown model parameters, whose values have to be estimated. These model parameters' values could be learnt from the observation (physical data) of the process that the model describes, denoted by z. Calibration is the process of finding the input parameter values that allow the computer model to be a trustworthy representation of the physical process; and is commonly used to solve the inverse problem [11,19,20]. Since the actual observations of the physical process of interest are considered, the *observation error* (measurement uncertainty) e should be included as part of the estimation process.

Another source of uncertainty, named *model discrepancy*, denoted as η, arises from the notion that the model is not a perfect representation of the true process due to the process complexity and the lack of computational resources. The *model discrepancy* is a complex concept to grasp, and we provide an example with climate models. We previously mentioned that these models solve coupled PDEs discretized over the vertical and horizontal grids over the sphere numerically to study the effect of input variables of interest on the climate [4]. For GCMs, grids with cell sizes of the order of 100–300 km horizontally are typically used, which leads to the inability to calculate the effect of clouds explicitly [21]. These processes are referred to as 'sub-grid scale processes', and their effect is approximated inside the model. As a result, *model discrepancy*, the difference between the physical phenomena and the model representation, will appear.

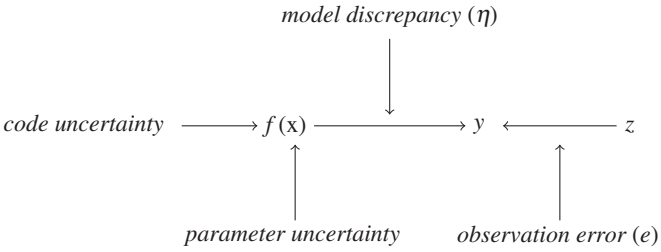

Figure 1. Schematic of the framework for analysing physical process y from computer model f and past physical data (observation) z and synthesizing all of the major uncertainties.

To perform model-based inference and to learn about the relationships between parameters $\mathbf{x}$ and model outputs $f(\mathbf{x})$, we are required to evaluate the model at many input parameter settings. In reality, f is computationally expensive to deal with, and therefore we treat f as uncertain at unseen $\mathbf{x}$. *Code uncertainty* represents our uncertainty about $f(\mathbf{x})$ at an arbitrary input value $\mathbf{x}$ [11]. An emulator conveniently provides a representation of the simulator behaviour together with the description of uncertainty about $f(\mathbf{x})$ across the input space (see §3a for more details).

Apart from calibration, mentioned previously, there are other forms of analysis that a modeller might be interested in performing. Uncertainty analysis deals with the parameter uncertainty by looking at the distribution of the computer model output induced by a probability distribution on input parameters [22]. History matching, an efficient global parameter search approach, is another uncertainty analysis tool, considered in detail in §3b. It uses emulation to find input parameter space regions for which the simulator output agrees with the observation and includes our uncertainty judgements [23]. Sensitivity analysis looks at identifying how model input parameters affect the model outputs [24]. The main effect index (first-order Sobol' index) and the total effect index (total Sobol' index) are commonly used measures of sensitivity of computer model output to an individual parameter x_i. The main effect index is based on considering the expected reduction in the uncertainty in the computer model output after we learn the true value of x_i [25]. The total effect index is based on quantifying the remaining uncertainty in the computer model output after we have learnt everything except x_i [26]. Sensitivity and uncertainty analysis are traditionally performed by employing a Monte Carlo algorithm. The algorithm proceeds as follows: the input parameter settings are drawn from the pre-defined probability distribution, the computer model runs are obtained at these parameter settings, which results in a random sample from the output distribution of interest. Kennedy & O'Hagan [11] pointed out that uncertainty and sensitivity analysis become impractical for expensive computer models and proposed to replace simulator with its surrogate. The emulator has been implemented as part of uncertainty and sensitivity analysis [22,24].

3. Methodologies in UQ

Here, we introduce a statistical model for the link between the computer model $f(\mathbf{x})$ and physical process y, which contains major sources of uncertainty, described in §2. The main components of this framework are emulation and history matching. We employ these tools to derive the input parameter settings that result in acceptable matches between model output and observed data. We demonstrate the use and performance of each individual component on a simple one-dimensional toy model.

(a) Emulation

In this paper, we consider a computer model as a black box model, i.e. the model is viewed in terms of its inputs $\mathbf{x}$ in p-dimensional parameter space $\mathcal{X}$ and outputs $f(\mathbf{x})$. Most models are complex and require a significant computational time to produce a single run. Model simulation

process could be a major bottleneck in model-based inference. Therefore, cheap approximation tools such as neural networks, splines and polynomial chaos could be used to approximate computer model behaviour across the input space [27–30]. Since we use an approximation, we should acknowledge and report our uncertainty about the computer model output (code). In this paper, we adopt a Gaussian process (GP) emulator as a surrogate to a complex computer model output. Contrary to other surrogates, GP emulator conveniently provides us with the measure of uncertainty about the generated prediction at an arbitrary input point $\mathbf{x}$, often expressed as variances. This measure of uncertainty corresponds to the code uncertainty defined in §1.

We consider an emulator as a sum of two processes [8], defined as

$$\left.\begin{aligned} f(\mathbf{x}) &= \sum_{j=1}^{q} \beta_j h_j(\mathbf{x}) + \epsilon(\mathbf{x}), \\ \epsilon(\mathbf{x}) &\sim GP\big(0, \sigma^2 r(\cdot,\cdot;\boldsymbol{\delta})\big), \end{aligned}\right\} \tag{3.1}$$

where $h(\mathbf{x})$ is a $(q \times 1)$ vector of specified regression functions in $\mathbf{x}$, and $\boldsymbol{\beta}$ is a vector of corresponding regression coefficients to be fitted. The second component of the model, $\epsilon(\mathbf{x})$, is modelled as a zero-mean GP; the $r(\cdot,\cdot;\boldsymbol{\delta})$ is pre-specified correlation function, and the $\boldsymbol{\delta}$ is vector of its parameters (correlation lengths), σ^2 corresponds to the variance parameter of the GP. We consider $h(\mathbf{x})^T\boldsymbol{\beta}$ as a global response surface, capturing dominant features of computer model output and $\epsilon(\mathbf{x})$ as a correlated residual process, depicting local input dependent deviation from the global response surface. We can represent the output of a computer model by a GP and proceed to derive the GP prior for $f(\mathbf{x})$ determined by a mean and covariance functions, i.e.

$$E[f(\mathbf{x})] = h(\mathbf{x})^T\boldsymbol{\beta}$$

and

$$\mathrm{Cov}[f(\mathbf{x}), f(\mathbf{x}')] = \sigma^2 r(\mathbf{x}, \mathbf{x}'; \boldsymbol{\delta}).$$

Using probabilistic notation, we define the probability distribution for $f(\mathbf{x})$ conditioned on the statistical model parameters $\{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}\}$ as

$$f(\mathbf{x})|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim GP\big(h(\mathbf{x})^T\boldsymbol{\beta}, \sigma^2 r(\cdot,\cdot;\boldsymbol{\delta})\big). \tag{3.2}$$

The regression functions in $h(\mathbf{x})$ can be anything from simple monomials to Fourier transformations of $\mathbf{x}$ [31], based on the expert opinion on the simulator behaviour or regression modelling (stepwise regression) [32,33]. Williamson *et al.* [31] stated that $h(\mathbf{x})$ could be used to add 'physical insights' into the statistical model. Meanwhile, the covariance function is used to characterize the similarity between the model output at two input points $\mathbf{x}$ and $\mathbf{x}'$ and explicitly depends on the form of the correlation function. Williams & Rasmussen [34] provided the description of a number of commonly used correlation functions in the computer experiments literature. For instance, the power exponential correlation function is defined as

$$r(\mathbf{x}, \mathbf{x}'; \boldsymbol{\delta}) = \exp\left\{ -\sum_{i=1}^{p} \left(\frac{x_i - x_i'}{\delta_i} \right)^{\phi_i} \right\},$$

where $\delta_i > 0$ and $0 < \phi_i \leq 2$. In this paper examples, we use a squared exponential correlation function with $\phi_i = 2, i = 1, \ldots, p$. The components of $\boldsymbol{\delta}$ are referred to as correlation lengths. We tend to obtain a stronger/weaker correlation for a pair of input points in the ith direction for larger/smaller values of the ith entry of the correlation length vector, i.e. δ_i.

To complete the construction of GP emulators, we are required to obtain an ensemble of runs of the computer model for updating equation (3.2). Suppose we observe n computer model realizations $\mathrm{F} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ at design points $\mathrm{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and let the ensemble be denoted

by {X, F}. By employing equation (3.2), the distribution of F is multivariate normal,

$$\mathrm{F}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim MVN(H\boldsymbol{\beta}, \sigma^2 K), \tag{3.3}$$

where $H = [h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)]^T$ is a $(q \times n)$ regression matrix, and K is $(n \times n)$ correlation matrix, with $(K)_{ij} = r(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\delta})$. The result of updating is the posterior distribution $f(\mathbf{x})|\{\mathrm{X}, \mathrm{F}\}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}$, i.e.

$$f(\mathbf{x})|\{\mathrm{X}, \mathrm{F}\}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim GP(m^*(\mathbf{x}), C^*(\cdot, \cdot)), \tag{3.4}$$

with mean

$$m^*(\mathbf{x}) = E_{\mathrm{F}}[f(\mathbf{x})] = h(\mathbf{x})^T\boldsymbol{\beta} + r(\mathbf{x}, \mathrm{X})K^{-1}(\mathrm{F} - H\boldsymbol{\beta})$$

and covariance

$$C^*(\mathbf{x}, \mathbf{x}') = Cov_{\mathrm{F}}[f(\mathbf{x}), f(\mathbf{x}')] = \sigma^2\Big[r(\mathbf{x}, \mathbf{x}') - r(\mathbf{x}, \mathrm{X})K^{-1}r(\mathrm{X}, \mathbf{x}')\Big],$$

where $r(\mathbf{x}, \mathrm{X})$ is n-vector whose ith component is $r(\mathbf{x}, \mathbf{x}_i), i = 1, \ldots, n$, the correlation between the point of interest $\mathbf{x}$ and the design point $\mathbf{x}_i$.

The GP model parameters are unknown and therefore have to be estimated or marginalized. [12] fixed values of GP hyperparameters at maximum likelihood estimates. Haylock & O'Haga [35] proposed to specify the non-informative prior for $\boldsymbol{\beta}$ and σ^2, i.e. $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ and carry out the marginalization. The final step is to estimate the values of correlation length parameters $\boldsymbol{\delta}$ either by finding the maximum likelihood estimate [11] or employing full Bayesian Markov chain Monte Caro (MCMC) methods. The probability distribution for f conditional on ensemble {X, F} and $\boldsymbol{\delta}$ is Student process with $n - p$ degrees of freedom. Oakley & O'Hagan [24] proposed to adopt a conjugate prior for $\boldsymbol{\beta}$ and σ^2 in order to incorporate prior beliefs about simulator into the model-based inference. Full Bayesian MCMC methods were adopted by [8,20,36]. These approaches account for the uncertainty in GP hyperparameters but come at a higher computational cost.

In this paper, we obtain MAP (maximum *a posteriori*) estimates for GP hyperparameters by optimizing the posterior distribution function with subjective priors. To derive these values, we used ExeterUQ_MOGP [37], an R interface for fitting GP emulators and performing UQ, based on mogp_emulator, a Python package developed by the Research Engineering Group at the Alan Turing Institute [38].

We produced a GP emulator for a one-dimensional toy function. The true function has the following form:

$$f(x) = \exp(0.5x) + 4\cos(x),$$

which has been computed at only 6 input points (training set X) evenly spread between $x_1 = -4$ and $x_6 = 3$. We specified the regression functions $h(x) = (1, x, x^2)^T$. In figure 2, the black solid and blue dashed lines correspond to $E_{\mathrm{F}}[f(\mathbf{x})]$ and $E_{\mathrm{F}}[f(\mathbf{x})] \pm 2\sqrt{\mathrm{Var}_{\mathrm{F}}[f(\mathbf{x})]}$. The true function $f(x)$ is given by a solid red line, and it can be seen that it lies within the prediction interval for all x, except values close to the boundaries of input space. We observe that the point predictor goes through the six points of the training set and our uncertainty drops down to zero at these points. However, as we move farther away from these points, the length of prediction interval increases, indicating that our uncertainty about model behaviour grows.

We presented GP emulator for a deterministic model, which produces the same output under the same model conditions. However, stochastic models are being used increasingly in epidemic models [39], engineering [40] and climate and weather models with stochastic parameterization [41,42] . The outputs of stochastic simulators possess inherent randomness. In this case, the presented emulation approach could be extended by including an independent noise term $\nu(\cdot)$. The log variance, $\log(\nu^2(\cdot))$, is another GP [39,43].

(b) Connecting model to reality

Emulation has two attractive features in our analysis. Firstly, emulators are extremely fast to evaluate, they can replace the original model in any larger calculation. Secondly, we consider the emulator as a probabilistic description about the value of the simulator at each input value, which

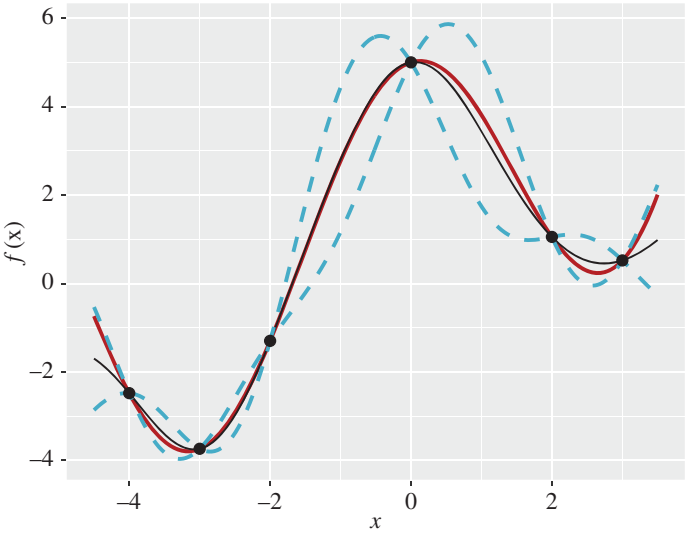

Figure 2. Plot of (true) function $f(x)$, $x \in [-4, 3]$ (red line). The black dots represent the observed data at 6 equally spaced values of x. The solid line represent the emulator's updated expectation $E_F[f(\mathbf{x})]$, and the pair of blue dashed lines give the credible interval $E_F[f(\mathbf{x})] \pm 2\sqrt{\mathrm{Var}_F[f(\mathbf{x})]}$, both as functions of x. (Online version in colour.)

we proceed to employ in the statistical model that establishes a relationship between observations of real system z and output of a model f. Kennedy & O'Hagan [11] introduced the 'best input approach' model that represents observation z via

$$z = y + e,$$

where y corresponds to the quantity of interest being observed (measured) and e corresponds to the error on this observation. These two quantities combined together give rise to the observation z. In this model specification, z, y and e represent random quantities. Observation error term e is assumed to be independent from y, and unbiased with zero mean and variance $\mathrm{Var}[e]$.

The 'best input approach' assumes the existence of a 'best input' setting $\mathbf{x}^*$ so that computer model represent the physical process within a specified statistical model

$$y = f(\mathbf{x}^*) + \eta,$$

where η is a model discrepancy term (structural error). This term indicates that model simulated at its 'best input', $f(\mathbf{x}^*)$, would still not be in agreement with the real physical system y. In particular, model discrepancy allows modellers to include any extra information about model's deficiencies in their model of real physical process [23]. We consider y, f, $\mathbf{x}^*$ and η as random quantities and specify zero expectation and variance $\mathrm{Var}[\eta]$ for η. The model discrepancy random quantity could be estimated based on expert judgements in combination with simulations from a more complex computer model [31].

By employing a statistical model described above together with emulation, we could perform a global parameter search. History matching attempts to find input parameter values to achieve the consistency between observations and computer model representation. The goal of history matching is to identify the regions of input space corresponding to acceptable matches, and this is performed by ruling out the implausible regions iteratively in waves. In particular, we are trying to rule out regions in $\mathcal{X}$ that could not contain $\mathbf{x}^*$ given the uncertainty specification and using the implausibility function, that has the following form:

$$\mathcal{I}(\mathbf{x}) = \frac{|z - E_F[f(\mathbf{x})]|}{\sqrt{\mathrm{Var}_F[z - E_F[f(\mathbf{x})]]}}. \tag{3.5}$$

We proceed to rewrite the denominator of the implausibility function as

$$\mathrm{Var}_{\mathrm{F}}[z - E_{\mathrm{F}}[f(\mathbf{x})]] = \mathrm{Var}[e] + \mathrm{Var}[\eta] + \mathrm{Var}_{\mathrm{F}}[f(\mathbf{x})].$$

To perform history matching, we require an emulator, so that we can obtain expectation, $E_{\mathrm{F}}[f(\mathbf{x})]$, and variance, $\mathrm{Var}_{\mathrm{F}}[f(\mathbf{x})]$, for any input setting $\mathbf{x}$. For model output f and observation z, large values of $\mathcal{I}(\mathbf{x})$ at any $\mathbf{x}$ imply that, relative to our uncertainty, the predicted output of computer model at $\mathbf{x}$ is very far from where we would expect it to be if $f(\mathbf{x})$ were consistent with z. However, small values of implausibility function imply either that we expect the model to be close to our observations for $\mathbf{x}$ or that we are very uncertain about the model behaviour, i.e. $\mathrm{Var}_{\mathrm{F}}[f(\mathbf{x})]$ is large. We are required to specify a value of a threshold, a, so that $\mathbf{x}$ at which $\mathcal{I}(\mathbf{x}) > a$ is deemed as implausible. For instance, [44] proposed to set the value of a to be 3 following 3 sigma rule. The remaining parameter space is termed as Not Ruled Out Yet (NROY) and defined as [16]

$$\mathcal{X}_1 = \mathcal{X}_{NROY} = \{\mathbf{x} \in \mathcal{X} : \mathcal{I}(\mathbf{x}) \leq a\}.$$

We could perform history matching iteratively. Refocussing is the process of performing history matching multiple times, by deriving wave k NROY space from the parameter space $\mathcal{X}_{k-1}$. We start by defining NROY space in wave k as

$$\mathcal{X}_k = \left\{\mathbf{x} \in \mathcal{X}_{k-1} : \mathcal{I}(\mathbf{x}, \mathrm{F}_{[k]} \leq a)\right\}.$$

We compute $\mathcal{I}(\mathbf{x}; \mathrm{F}_{[k]})$ across the $\mathcal{X}_{k-1}$ using expectation and variance produced by an emulator for $f(\mathbf{x})$ defined inside $\mathcal{X}_{k-1}$. To construct this emulator, we are required to obtain the design

$$\mathrm{X}_{[k]} = (\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,n_k})^T \in \mathcal{X}_{k-1},$$

together with the corresponding computer model simulations, i.e.

$$\mathrm{F}_{[k]} = (f(\mathbf{x}_{k,1}), \ldots, f(\mathbf{x}_{k,n_k}))^T.$$

We retain $\mathbf{x}$ as part of wave k NROY space only if it has not been ruled out (RO) in the previous $k-1$ waves of history matching.

Refocussing is considered as a powerful method and has been applied across the wide range of fields including galaxy formation [2,45], HIV transmission model [46,47] and climate [31,48]. At each iteration (wave), we increase the density of ensemble, which leads to the improvement in the performance of statistical emulators, i.e. more accurate predictions and lower uncertainty about the predictions. Therefore, we expect to retain as part of NROY space input parameters setting at which model output is close to the observation.

We perform history matching on a simple one-dimensional toy model introduced in §3a. Figure 3 depicts the one-dimensional toy model with both observation error e and model discrepancy η. Here, we show the model behaviour $f(x)$, given by the red solid line, together with model error (the red dashed lines represent $f(x) \pm 2\sqrt{\mathrm{Var}[\eta]}$). The observation z is given by the solid grey line together with the observation error (the grey dashed lines show $z \pm 2\sqrt{\mathrm{Var}[e]}$). We specify the following values for our demonstration: $z = 4.75$, $\mathrm{Var}[e] = 0.1$ and $\mathrm{Var}[\eta] = 0.1$

In figure 4*a*, we demonstrate the emulator expectation and credible intervals as in figure 2 together with the observation plus observed error. The colours on the x-axis correspond to the implausibility values $\mathcal{I}(x)$: red and green for RO ($\mathcal{I}(x) > 3$) and NROY ($\mathcal{I}(x) < 3$), respectively.

The function $f(x)$ is defined on $-4.5 < x < 3.5$. We proceed to perform history matching and specify the threshold a at 3 to obtain the wave 1 NROY space $\mathcal{X}_1$, shown by the green region on the x-axis in figure 4*a*, i.e. $-1.16 < x < 1.37$. We perform the second wave of history matching by obtaining an additional model run within the $\mathcal{X}_1$ space, reconstructing the emulator, and recalculating the implausibility measure $\mathcal{I}(x)$. In figure 4*b*, we observe that the emulator has become more accurate together with reduced uncertainty about its predictions. As a result, the non-implausible region in green has shrunk, i.e. $-0.96 < x < 1.13$. In case if we ignored observation error, model discrepancy and code uncertainty in our analysis, only two values of x can be viewed as acceptable while all others are unacceptable.

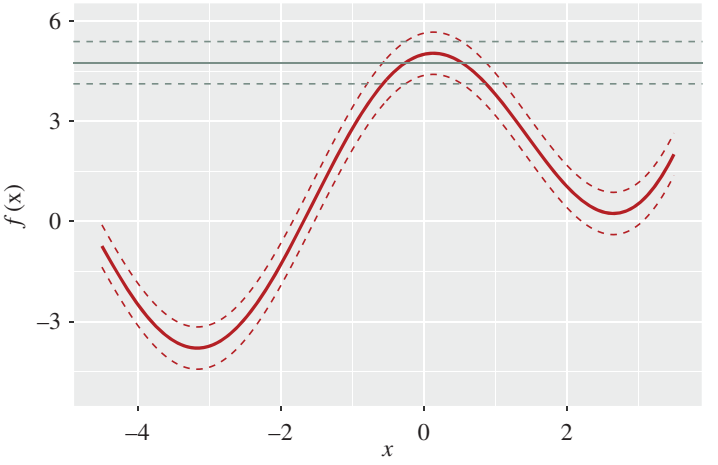

Figure 3. The model $f(x)$ is given by the red line, the observed data z by the horizontal grey line. We include both observation error e (the grey dashed lines represent $z \pm 2\sqrt{\text{Var}[e]}$) and model discrepancy η (the red dashed lines show $f(\mathbf{x}) \pm 2\sqrt{Var[\eta]}$). (Online version in colour.)

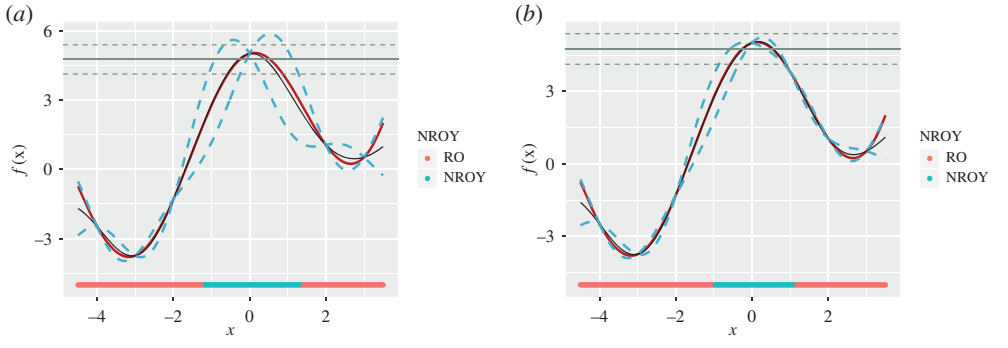

Figure 4. (*a*) The emulator expectation and credible intervals as in figure 2; however, now the observation z plus observed error has been included as the horizontal grey solid and dashed lines respectively. The implausibilities $\mathcal{I}(x)$ are represented by the colours on the x-axis: red and green for high ($\mathcal{I}(x) > 3$) and low ($\mathcal{I}(x) < 3$) implausibility respectively, with the green interval defining the non-implausible region $\mathcal{X}_1$. (*b*) The second wave is performed by evaluating an additional point located within $\mathcal{X}_1$. The emulator becomes more accurate over $\mathcal{X}_1$ and the implausibility more strict, hence defining the smaller non-implausible region $\mathcal{X}_2$, given by the green interval. (Online version in colour.)

4. Example

We proceed to demonstrate the use of emulation and history matching in the climate model context. In particular, this example is inspired by the work that has been done as part of the ANR (Agence Nationale de la Recherche) funded HIGH-TUNE project. The main objective of HIGH-TUNE project is to improve the representation of the boundary-layer clouds in two French General Circulation Models, ARPEGE-Climat and LMDZ. These types of clouds play a crucial role in the water and energy cycles of the atmosphere and impact surface temperature at various scales [49]. However, boundary layer clouds are much smaller than a grid cell of a climate model, and therefore the collective behaviour and the effect on the large-scale model outputs of an ensemble of boundary-layer clouds is parameterized. The parameterization schemes depend on a variety of 'free' parameters and calibrating (tuning) these parameters is crucial to avoid biases in the global climate models [4].

In particular, climate modellers are interested in finding the regions of input parameters, that correspond to acceptable matches between single-column simulations of the GCMs (SCMs) and three-dimensional high-resolution large eddy simulations (LESs). Note that LES runs are treated as surrogate observations since it is challenging to collect data on clouds' properties at the required temporal and spatial scale [50]. In climate modelling, this type of tuning is termed 'process-based tuning'.

A number of approaches could be used to perform climate model tuning. The uncertain parameters could be adjusted manually, which is referred to as expert tuning [51–54]. However, this approach suffers from a lack of objectivity as well as being very time-consuming and often requires thousands of runs of a climate model. Another approach is based on specifying and optimizing a cost function with respect to calibration parameters that measures the distance between model simulations and a collection of observations [55,56]. However, constantly evaluating climate models at new parameter settings is time-consuming, especially as the complexity of the model increases. The solution to this problem could be to construct surrogate models (emulators) for target metrics and use surrogate models' outputs in the cost function computation [51,57]. However, this framework fails to account for a number of important sources of uncertainty.

We proceed to present the application of Bayesian emulation and history matching to climate model tuning. For demonstration, we consider the statistics on water vapour at 500 m (`qv500`) generated by the SCM by varying five input parameters, that are part of convection parameterization, for two cases SANDU [58] and ARMCU [59]. These two cases correspond to the specific characteristics of the physical representation of boundary-layer clouds. This metric for two cases has been averaged over a few hours, for more details see [60][Table 2]. The experiment aimed to derive the acceptable matches between SCM's output and LES for these two cases, and we treated them as two independent outputs, i.e. $f_i(\mathbf{x}),\ i = 1, 2$.

First, we produced a space-filling design X over the whole input space $\mathcal{X}$, i.e. a 90-point maximin Latin hypercube design [61]. We evaluated model at the obtained design for each case. To construct a GP emulator, we were required to specify the form of the regression function, $h(\mathbf{x})$, and the prior distributions for GP parameters, i.e. $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, σ^2. We obtained the form of a mean function by using a stepwise regression procedure [62]. We used a forwards and backwards stepwise selection routine and considered interaction terms, higher-order polynomials together with the Fourier functions for selection. Similar procedure has been used by [33]. We used the default prior specification setting implemented as part of Exeter_MOGP [37]. In particular, we specified a uniform prior for intercept and $\mathrm{N}(0, 10^2)$ for the regression coefficients. These weakly informative priors rule out unreasonable parameter values but are not so strong to rule out values that might make sense in the context of model data. A logNormal(0, 0.125) was defined for correlation length parameter. We specified a subjective prior for the variance parameter σ^2, so that together with the prior specifications for $\boldsymbol{\delta}$ the confounding between σ^2, regression line and $\boldsymbol{\delta}$ could be resolved. Finally, MAP estimates were obtained for GP hyperparameters.

Prior to history matching, we ran diagnostic checks to assess the performance of the obtained emulators. In particular, Leave-One-Out cross-validation was performed, where each point from the design set was retained for validation, while the emulator was refitted. Figure 5 shows the diagnostic plots for four of the model parameters for SANDU (top row) and ARMCU (bottom row) cases. Black points and error bars (±2 s.d. prediction intervals) are computed from $E_{\mathrm{F}}[f(\mathbf{x})]$ and $Var_{\mathrm{F}}[f(\mathbf{x})]$. The true (left out) values are plotted in green/red if they lie within/outside two standard deviation prediction intervals.

The plots in the second row indicate that the emulator represents model well for ARMCU case with small error bars and predictions close to true values. In the top row of figure 5, we observe four points outside the uncertainty bounds. Since 2 standard deviations correspond to 95% confidence intervals, we expect 5% of points to be red and outside the error bars. We also obtain larger error bars due to more sophisticated behaviour of model for SANDU case. Overall, we are happy with the emulators' representation of the model outputs and proceed to perform history matching.

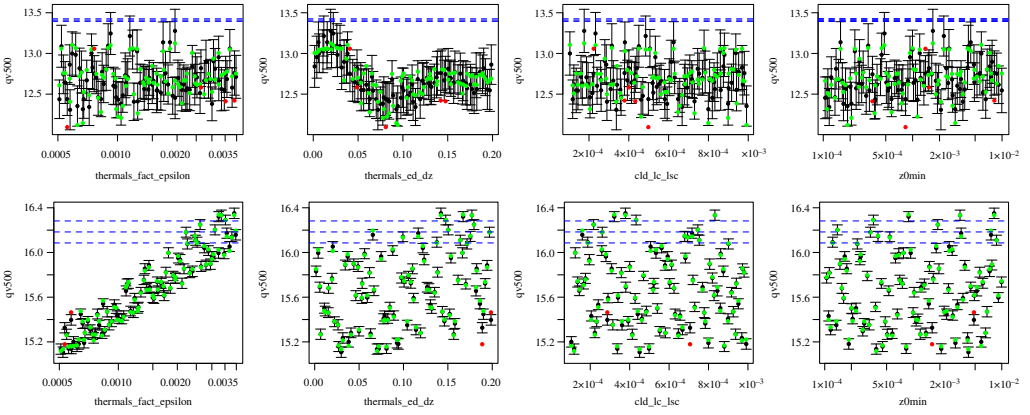

Figure 5. Leave-One-Out diagnostics plots against each of the parameters for SANDU (top row) and ARMCU (second row) cases on original input scales. The predictions and two standard deviation prediction intervals are in black. The true model values are in green if they lie within two standard deviation prediction intervals, or red otherwise. The observation z plus observed error ($z \pm 2\sqrt{\mathrm{Var}[e]}$) are shown by blue dashed lines. (Online version in colour.)

Table 1. Summary information for history matching.

case	observation z	observation error Var[e]	model discrepancy Var[η]
SANDU	13.41	0.0078	0.040
ARMCU	16.18	0.049	0.001

We provide the specification of z, Var[e] and Var[η] for each case in table 1. We imposed the implausibility constraint simultaneously across two outputs, by employing $\mathcal{I}_M < c$ with conservative cutoff c of 3, where $\mathcal{I}_M(\mathbf{x})$ is the maximum implausibility measure defined by $\mathcal{I}_M(\mathbf{x}) = \max_i \mathcal{I}_i(\mathbf{x})$, $i = 1, 2$. The alternative measure is to consider the second implausibility, which allows for some inaccuracy of the emulators. We performed a single wave of history matching and obtained $\mathcal{X}_1$, the NROY space remaining after wave 1, which has a size of 28% of the original input space $\mathcal{X}$.

We proceeded to investigate two-dimensional representations of the shape of non-implausible region $\mathcal{X}_1$ for a selection of input parameters. The NROY density and minimum implausibility plots are shown in figure 6. These plots are produced by computing the implausibility function, given in equation (3.5), at a large number of points within the five-dimensional input space. Each panel on the upper triangle corresponds to the density of points in the NROY space. Grey regions are completely ruled out, which indicates that for a fixed value of two parameters under consideration, we are unable to retain any parameter settings in the other three dimensions. Each panel on the lower triangle shows the minimum implausibility plot. The value behind each pixel corresponds to the smallest implausibility found at the fixed value of a pair of parameters. The red regions indicate high implausibility, and we do not expect to observe good matches between climate model and data in these regions of the input space. Green/yellow regions correspond to the low implausibility values and the location of potentially 'good' input parameters settings. In general, we are interested in investigating further these regions in the subsequent waves of history matching.

Figure 6 provides modellers with insights into the relationships between a pair of input parameters. In particular, we observe that there is a non-linear relationship between parameters `thermals_fact_epsilon` and `thermals_ed_dz`, i.e. these parameters should be varied jointly to obtain NROY models. There is a very limited effect from `cld_lc_lsc` in our analysis. By performing a single wave of history matching, we managed to rule out more than 70% of input

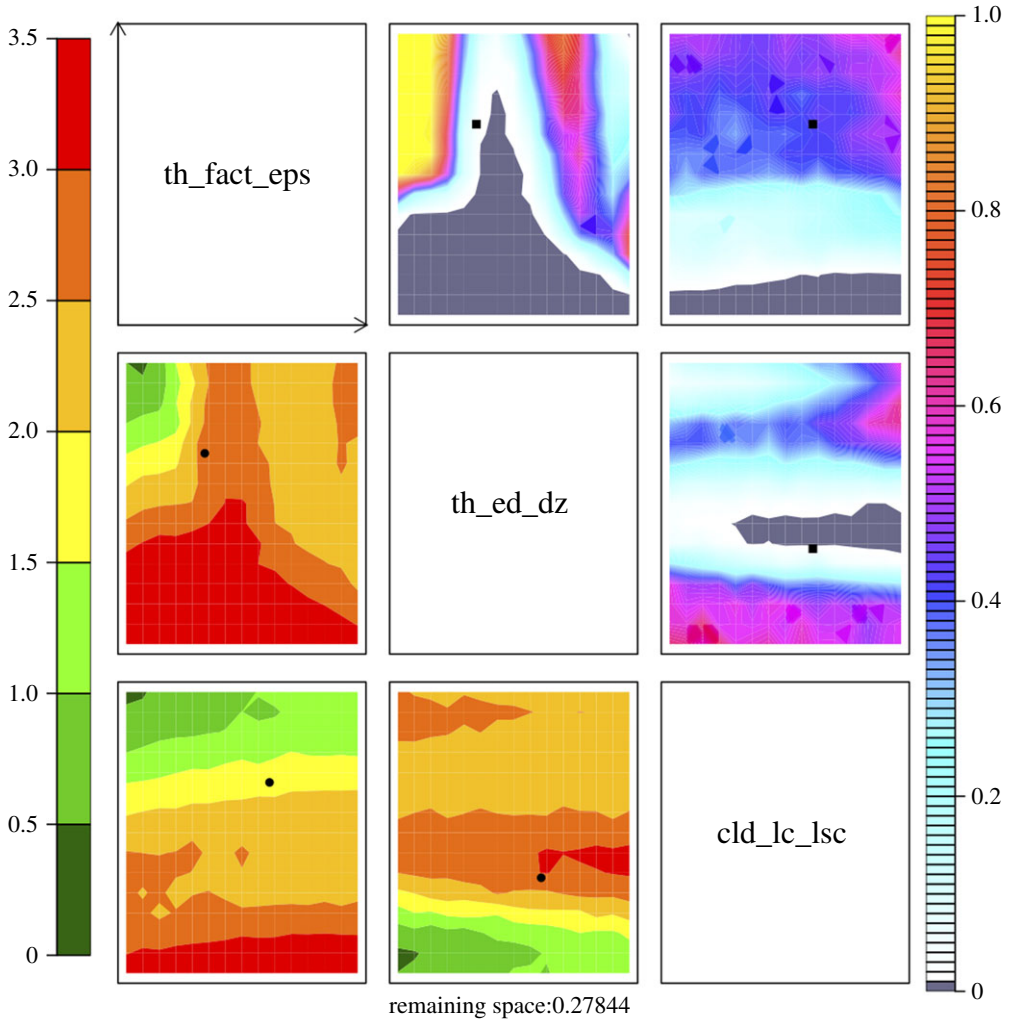

Figure 6. NROY density plots (upper triangle) and minimum implausibility plots (lower triangle). Each panel plots either NROY density or minimum implausibility for a pair of parameters. NROY densities, for each pixel on any panel in the upper triangle, represent the proportion of points in the input space behind that pixel that are NROY and are indicated by the colour whose scale is indicated on the right. Grey coloured regions are completely ruled out. Minimum implausibilities, for each pixel on any panel on the lower triangle of the picture, represent the smallest implausibilities found in input space. These plots are oriented the same way as those on the upper triangle, for the ease of visual comparison. Currently used parameter values in GCM is depicted as the square on the NROY density plots and as the circular point on the minimum implausibility plots. (Online version in colour.)

space. We note that the default values for the selected parameters shown in figure 6 lie inside the NROY space. These values were obtained by performing a slow expert tuning [60]. The second wave of history matching could be performed starting with obtaining climate model runs in the green/yellow regions and updating GP emulators.

5. Discussion

Increasingly science (and policy) is relying on numerical models of the physical world. The credibility of such models depends in part on the reproducibility of computer experiments. A large part of reproducibility comes from having well maintained and open code [63]. In this paper, we argue that reproducibility in computer models is not just about having the correct

code, important though that is, it is about the relationship between the inputs and outputs of the model as well as its relationship to data from the real world. Any results need to be presented with uncertainty bounds of some form and all assumptions (priors, etc.) need to be clearly and openly stated.

We have presented approaches that derive regions in input space at which we expect to achieve consistency between the computer model and observations of the system. In particular, the use of emulators and history matching allows modellers to explicitly include major sources of uncertainty discussed in §2. Simulating model at the obtained input settings results in a range of predictions, that we argue should be reported and analysed to ensure the reproducibility criteria are met.

We believe that the presented statistical analysis is important for model development. Brynjarsdóttir & O'Hagan [64] pointed out that modelling is an iterative process, i.e. by comparing the model to the observed data modellers could learn about model deficiencies to work on for the next release. Stainforth *et al.* [10] argued that uncertainty assessment and model development are part of the 'two-pronged approach'.

We recognize that the process of identifying sources of uncertainty is subjective and model dependent. For instance, there is still ongoing research in formulating model discrepancy, since the poor representation of this term leads to biased and over-confident input parameter values that cannot be used to generate trustworthy model predictions [64]. To specify model discrepancy Var[η], modellers' knowledge and opinions could be extracted via prior elicitation [65] in combination with simulations produced by a high-resolution model [10,31].

We need to mention an alternative probabilistic approach to history matching, a Bayesian calibration. Kennedy & O'Hagan [11] obtained a posterior distribution for input parameters of interest (calibration parameters), i.e. $\pi(\mathbf{x}^*|\{\mathrm{X}, \mathrm{F}, z\})$, which takes into account major sources of uncertainty. However, we choose to employ history matching in our study of complex computer models because of its flexibility to specify the different form of implausibility functions as well as the set of metrics of interest [2,48]. Another reason is that the result of Bayesian calibration is always the probability distribution for $\mathbf{x}^*$ over the input space $\mathcal{X}$. If the computer model cannot represent the system, then the calibration distribution is unfit to be employed in the further analysis. By contrast, history matching would rule out the entire input parameter space, which indicates that the computer model is unacceptable to represent the true physical system.

In this paper, we performed an analysis based on a single model. However, in climate science, it is common to perform inference about the physical process based on multiple models, i.e. 'ensemble' of different simulators or 'multi-model ensemble (MME)'. For this type of problem, [66] proposed to take into account the shared discrepancies among the simulators. In particular, the statistical model (mimic) is specified to represent the common structure in the behaviour of simulators. Therefore, these simulators are characterized by the parameters of a statistical model (descriptor). To model the relationship between the simulator and the real climate, [66] assumed that the simulator descriptor is centred upon $\theta_0 + \omega$, where θ_0 is a real climate descriptor and ω is an ensemble-specific discrepancy.

Data accessibility. This article has no additional data.

Authors' contributions. V.V. drafted the manuscript. All authors read and approved the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

1. Goldstein M. 2006 Subjective Bayesian analysis: principles and practice. *Bayesian Anal.* **1**, 403–420. (doi:10.1214/06-BA116)
2. Vernon I, Goldstein M, Bower R. 2014 Galaxy formation: Bayesian history matching for the observable universe. *Stat. Sci.* **29**, 81–90. (doi:10.1214/12-STS412)
3. Gettelman A, Rood RB. 2016 Demystifying climate models. *A Users Guide to Earth System Models*.

4. Hourdin F *et al.* 2017 The art and science of climate model tuning. *Bull. Am. Meteorol. Soc.* **98**, 589–602. (doi:10.1175/BAMS-D-15-00135.1)
5. WCRP. Coupled Model Intercomparison Project (CMIP). https://www.wcrp-climate.org/wgcm-cmip, 2020. (accessed 22 May 2020).
6. Treasury HM. 2015 The aqua book: guidance on producing quality analysis for government. *HM Government, London, UK, nd https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-qualityanalysis-for-government (accessed 10 July 2017).*
7. Treasury HM, McPherson N. 2013 Review of quality assurance of government analytical models: Final report. *Nick Macpherson, March*.
8. Williamson D, Blaker AT. 2014 Evolving Bayesian emulators for structured chaotic time series, with application to large climate models. *SIAM/ASA J. Uncertain. Quantif.* **2**, 1–28. (doi:10.1137/120900915)
9. Economist Editorial. The hard choices COVID policymakers face. *The Economist*, April 2020.
10. Stainforth DA, Allen MR, Tredger ER, Smith LA. 2007 Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. A* **365**, 2145–2161. (doi:10.1098/rsta.2007.2074)
11. Kennedy MC, O'Hagan A. 2001 Bayesian calibration of computer models. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **63**, 425–464. (doi:10.1111/1467-9868.00294)
12. Currin C, Mitchell T, Morris M, Ylvisaker D. 1991 Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Am. Stat. Assoc.* **86**, 953–963. (doi:10.1080/01621459.1991.10475138)
13. Sacks J, Schiller SB, Welch WJ. 1989 Designs for computer experiments. *Technometrics* **31**, 41–47. (doi:10.1080/00401706.1989.10488474)
14. Santner TJ, Williams BJ, Notz WI, Williams BJ. 2003 *The design and analysis of computer experiments*, vol. 1. New York, NY: Springer.
15. Craig PS, Goldstein M, Seheult AH, Smith JA. 1996 Bayes linear strategies for matching hydrocarbon reservoir history. *Bayesian Stat.* **5**, 69–95.
16. Williamson D, Blaker AT, Hampton C, Salter J. 2015 Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.* **45**, 1299–1324. (doi:10.1007/s00382-014-2378-z)
17. Der Kiureghian A, Ditlevsen O. 2009 Aleatory or epistemic? Does it matter? *Struct. Saf.* **31**, 105–112. (doi:10.1016/j.strusafe.2008.06.020)
18. Werndl C. 2019 Initial-condition dependence and initial-condition uncertainty in climate science. *Br. J. Phil. Sci.* **70**, 953–976. (doi:10.1093/bjps/axy021)
19. Chang K-L, Guillas S. 2019 Computer model calibration with large non-stationary spatial outputs: application to the calibration of a climate model. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **68**, 51–78. (doi:10.1111/rssc.12309)
20. Higdon D, Gattiker J, Williams B, Rightley M. 2008 Computer model calibration using high-dimensional output. *J. Am. Stat. Assoc.* **103**, 570–583. (doi:10.1198/016214507000000888)
21. Diallo FB, Hourdin F, Rio C, Traore A-K, Mellul L, Guichard F, Kergoat L. 2017 The surface energy budget computed at the grid-scale of a climate model challenged by station data in West Africa. *J. Adv. Model. Earth Syst.* **9**, 2710–2738. (doi:10.1002/2017MS001081)
22. Oakley J, O'Hagan A. 2002 Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**, 769–784. (doi:10.1093/biomet/89.4.769)
23. Vernon I, Liu J, Goldstein M, Rowe J, Topping J, Lindsey K. 2018 Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC Syst. Biol.* **12**, 1. (doi:10.1186/s12918-017-0484-3)
24. Oakley JE, O'Hagan A. 2004 Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Stat. Soc. B (Stat. Methodol.)* **66**, 751–769. (doi:10.1111/j.1467-9868.2004.05304.x)
25. Saltelli A, Chan K, Scott EM. 2000 Wiley series in probability and statistics. In *Sensitivity Analysis*. Wiley.
26. Homma T, Saltelli A. 1996 Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **52**, 1–17. (doi:10.1016/0951-8320(96)00002-6)
27. Chen VCP, Tsui KL, Barton RR, Meckesheimer M. 2006 A review on design, modeling and applications of computer experiments. *IIE Trans.* **38**, 273–291. (doi:10.1080/07408170500232495)

28. Jin R, Chen W, Simpson TW. 2001 Comparative studies of metamodelling techniques under multiple modelling criteria. *Struct. Multidiscip. Optim.* **23**, 1–13. (doi:10.1007/s00158-001-0160-4)
29. Mohammadi H, Challenor P, Goodfellow M. 2019 Emulating dynamic non-linear simulators using Gaussian processes. *Comput. Stat. Data Anal.* **139**, 178–196. (doi:10.1016/j.csda.2019.05.006)
30. Owen NE, Challenor P, Menon PP, Bennani S. 2017 Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators. *SIAM/ASA J. Uncertain. Quantif.* **5**, 403–435. (doi:10.1137/15M1046812)
31. Williamson DB, Blaker AT, Sinha B. 2017 Tuning without over-tuning: parametric uncertainty quantification for the nemo ocean model. *Geosci. Model Dev.* **10**, 1789–1816. (doi:10.5194/gmd-10-1789-2017)
32. Cumming JA, Goldstein M. 2009 Small sample Bayesian designs for complex high-dimensional models based on information gained using fast approximations. *Technometrics* **51**, 377–388. (doi:10.1198/TECH.2009.08015)
33. Williamson D, Goldstein M, Allison L, Blaker A, Challenor P, Jackson L, Yamazaki K. 2013 History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.* **41**, 1703–1729. (doi:10.1007/s00382-013-1896-4)
34. Williams CKI, Rasmussen CE. 2006 *Gaussian processes for machine learning*, vol. 2. Cambridge, MA: MIT press.
35. Haylock RG, O'Hagan A. 1996 On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. *Bayesian Stat.* **5**, 629–637.
36. Kaufman CG, Sain SR. 2010 Bayesian functional {ANOVA} modeling using Gaussian process prior distributions. *Bayesian Anal.* **5**, 123–149. (doi:10.1214/10-BA505)
37. University of Exeter. ExeterUQ_MOGP. https://github.com/BayesExeter/ExeterUQ_MOGP, 2020. (accessed 4 June 2020).
38. Alan Turing Institute. mogp_emulator. https://github.com/alan-turing-institute/mogp_emulator, 2020. (accessed 4 June 2020).
39. Binois M, Huang J, Gramacy RB, Ludkovski M. 2019 Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics* **61**, 7–23. (doi:10.1080/00401706.2018.1469433)
40. Ankenman B, Nelson BL, Staum J. 2008 Stochastic kriging for simulation metamodeling. In *2008 Winter Simulation Conference*, pp. 362–370. IEEE.
41. Leutbecher M *et al.* 2017 Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Q. J. R. Meteorol. Soc.* **143**, 2315–2339. (doi:10.1002/qj.3094)
42. Palmer TN. 2012 Towards the probabilistic earth-system simulator: a vision for the future of climate and weather prediction. *Q. J. R. Meteorol. Soc.* **138**, 841–861. (doi:10.1002/qj.1923)
43. Baker E, Challenor P, Eames M. 2020 Predicting the output from a stochastic computer model when a deterministic approximation is available. *J. Comput. Graph. Stat.* **29**, 1–12.
44. Pukelsheim F. 1994 The three sigma rule. *Am. Stat.* **48**, 88–91.
45. Bower RG, Vernon I, Goldstein M, Benson AJ, Lacey CG, Baugh CM, Cole S, Frenk CS. 2010 The parameter space of galaxy formation. *Mon. Not. R. Astron. Soc.* **407**, 2017–2045. (doi:10.1111/j.1365-2966.2010.16991.x)
46. Andrianakis I, McCreesh N, Vernon I, McKinley TJ, Oakley JE, Nsubuga RN, Goldstein M, White RG. 2017 Efficient history matching of a high dimensional individual-based HIV transmission model. *SIAM/ASA J. Uncertain. Quantif.* **5**, 694–719. (doi:10.1137/16M1093008)
47. Andrianakis I, Vernon IR, McCreesh N, McKinley TJ, Oakley JE, Nsubuga RN, Goldstein M, White RG. 2015 Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda. *PLoS Comput. Biol.* **11**, e1003968. (doi:10.1371/journal.pcbi.1003968)
48. Salter JM, Williamson DB, Scinocca J, Kharin V. 2019 Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *J. Am. Stat. Assoc.* **114**, 1–24. (doi:10.1080/01621459.2018.1514306)
49. Bony S, Dufresne J-L. 2005 Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.* **32**, L20806. (doi:10.1029/2005GL023851)

50. Hourdin F *et al.* 2013 LMDZ5B: the atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection. *Clim. Dyn.* **40**, 2193–2222. (doi:10.1007/s00382-012-1343-y)
51. Bellprat O, Kotlarski S, Lüthi D, Schär C. 2012 Objective calibration of regional climate models. *J. Geophys. Res.: Atmospheres* **117**, D23115. (doi:10.1029/2012JD018262)
52. Gent PR *et al.* 2011 The community climate system model version 4. *J. Clim.* **24**, 4973–4991. (doi:10.1175/2011JCLI4083.1)
53. Mauritsen T *et al.* 2012 Tuning the climate of a global model. *J. Adv. Model. Earth Syst.* **4**, 484–502. (doi:10.1029/2012MS000154)
54. Watanabe M *et al.* 2010 Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. *J. Clim.* **23**, 6312–6335. (doi:10.1175/2010JCLI3679.1)
55. Zhang T, Li L, Lin Y, Xue W, Xie F, Xu H, Huang X. 2015 An automatic and effective parameter optimization method for model tuning. *Geosci. Model Dev.* **8**, 3579–3591. (doi:10.5194/gmd-8-3579-2015)
56. Zou L, Qian Y, Zhou T, Yang B. 2014 Parameter tuning and calibration of RegCM3 with MIT–Emanuel cumulus parameterization scheme over CORDEX East Asia domain. *J. Clim.* **27**, 7687–7701. (doi:10.1175/JCLI-D-14-00229.1)
57. Yang B *et al.* 2013 Uncertainty quantification and parameter tuning in the CAM5 Zhang-McFarlane convection scheme and impact of improved convection on the global circulation and climate. *J. Geophys. Res.: Atmos.* **118**, 395–415. (doi:10.1029/2012JD018213)
58. Sandu I, Stevens B. 2011 On the factors modulating the stratocumulus to cumulus transitions. *J. Atmos. Sci.* **68**, 1865–1881. (doi:10.1175/2011JAS3614.1)
59. Brown AR *et al.* 2002 Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Q. J. R. Meteorol. Soc.* **128**, 1075–1093. (doi:10.1256/003590002320373210)
60. Hourdin F *et al.* In press Process-based climate model development harnessing machine learning: II. Model calibration from single column to global. *J. Adv. Model. Earth Syst.* (doi:10.1029/2020MS002225)
61. McKay MD, Beckman RJ, Conover WJ. 1979 Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245.
62. Draper NR, Smith H. 1998 *Applied Regression Analysis*, vol. 326. New York, NY: Wiley.
63. Alan Turing Institute. The Turing Way - a handbook for reproducible data science. https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science, 2020. (accessed 4 June 2020).
64. Brynjarsdóttir J, O'Hagan A. 2014 Learning about physical parameters: the importance of model discrepancy. *Inverse Prob.* **30**, 114007. (doi:10.1088/0266-5611/30/11/114007)
65. O'Hagan A. 1998 Eliciting expert beliefs in substantial practical applications: [read before the Royal Statistical Society at meeting on 'elicitation' on wednesday, 16th April 1997, the president, professor AFM Smithin the chair]. *J. R. Stat. Soc.: D (The Statistician)* **47**, 21–35. (doi:10.1111/1467-9884.00114)
66. Chandler RE. 2013 Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Phil. Trans. R. Soc. A* **371**, 20120388. (doi:10.1098/rsta.2012.0388)