Available online at www.sciencedirect.com

ScienceDirect

Communications in Nonlinear Science and Numerical Simulation 00 (2015) 1–14

CNSNS

Bayesian uncertainty analysis for complex physical systems modelled by computer simulators with applications to tipping points

C.C.S. Caiado[a], M. Goldstein[a]

[a]*Department of Mathematical Sciences, University of Durham, United Kingdom*

Abstract

In this paper we present and illustrate basic Bayesian techniques for the uncertainty analysis of complex physical systems modelled by computer simulators. We focus on emulation and history matching and also discuss the treatment of observational errors and structural discrepancies in time series. We exemplify such methods using a four-box model for the termohaline circulation. We show how these methods may be applied to systems containing tipping points and how to treat possible discontinuities using multiple emulators.

Keywords: Bayesian analysis, emulation, history matching, uncertainty quantification, tipping points

1. Introduction

It is common to study complex physical systems using mathematical models, implemented as high dimensional computer simulators. In this paper, we describe certain characteristic features of this growing field of study which aims to quantify and synthesise all of the uncertainties involved in relating such simulators to physical systems, within the framework of Bayesian statistics. In particular, we focus on emulation and history matching as key tools to understand a complex simulator, and to investigate the underlying physical system represented. Our account is in two parts. The first part introduces and describes these tools, illustrating with application to the four-box model of the Atlantic presented in [24]. Most of the applications of emulation methodology assume smoothness of the system's response as a function of the input parameters to the simulator. In the second part of this account, we show how the ideas of emulation and history matching may be extended to deal with problems involving the kinds of discontinuity characteristic of simulators exhibiting tipping point behaviour, and again, illustrate this approach in the context of the Zickfield *et al.* model.

2. Computer simulators and physical systems

We are concerned with the use of computer simulators to reduce uncertainty about real world physical systems. A simulator f is a computer implementation of a, typically deterministic, complex computer model for a physical system. We denote the simulator as $f(\mathbf{x})$, where $\mathbf{x}$ are uncertain model parameters, corresponding to unknown system properties, and $f(\mathbf{x})$ is a vector of simulator outputs corresponding to system behaviour. We denote actual system behaviour as $\mathbf{y}$. In the problems we shall be concerned with below, $\mathbf{y}$ and thus $f(\mathbf{x})$ will typically be one, or more, time series.

We have n evaluations of the simulator at inputs $\mathbf{x} = (x_1, \ldots, x_n)$. We denote the resulting evaluations as $\mathbf{F} = (f(x_1), \ldots, f(x_n))$.

In order to relate the simulator and the physical system, we suppose that there is an appropriate value $\mathbf{x}^*$, for the system properties. For a perfect model, this would imply that $y = f(\mathbf{x}^*)$. However, in practice, f inevitably simplifies the system physics and usually also approximates the solution of the physical equations underlying the model. Therefore we suppose that

$$\mathbf{y} = f(\mathbf{x}^*) + \epsilon \tag{1}$$

where ϵ is a correlated random vector corresponding to the structural discrepancy between the simulator and the physical system, expressing the uncertainty about the physical system that would remain if we knew the evaluation of the simulator at $\mathbf{x}^*$. It is commonly assumed that ϵ is independent of f and $\mathbf{x}^*$. For now, we make this assumption also, but we shall modify it in one key respect in later sections. Here, and onwards, all probabilistic statements are understood in the Bayesian sense as uncertainty judgements of the analyst [18].

In our development, often we will not take the view that there is a unique appropriate value for x^*. Because of the imperfect way in which the model translates an approximate representation of system properties into a simplified version of system behaviour, we are often reluctant to make the judgement that one of the approximate representations of system properties should be considered as true. Instead we may prefer to identify the class of all such representations which are consistent with historical data. If we wish to carry out a further stage of Bayesian analysis to produce a posterior distribution over this class, then, of course, we are free to do so.

We further suppose that we have observations $\mathbf{z}$ on a historical subvector $\mathbf{y}_h$ of $\mathbf{y}$, corresponding to output subvector $f_h(\mathbf{x}^*)$. We suppose that $\mathbf{z}$ is related to $\mathbf{y}_h$ as

$$\mathbf{z} = \mathbf{y}_h + \mathbf{e} \tag{2}$$

where $\mathbf{e}$ is the measurement error, taken to be independent of $\mathbf{y}_h$.

We now introduce the example that we will use to illustrate the methods described in this paper.

3. Example: The Zickfeld et al (2004) model of the Atlantic

Highly complex systems are often represented using multi-compartment models where the distribution of the contents in each compartment is assumed homogeneous. In box climate models, it is often assumed that each compartment or box has a constant volume and homogeneous temperature; these models are often represented by a system of differential equations which make the problem tractable analytically and numerically. Box models are computationally cheap in comparison to global circulation models which discretize the diffusion of fluids over a fine grid in order to numerically approximate solutions to the Navier-Stokes equations. Moreover, the complexity of global models often masks the underlying dynamics of the system, obscuring information essential to draw forecasts and inferences regarding the equilibrium of the system.

Here we look at a four-box model of the Atlantic proposed by Zickfield *et al.* [24] as an extension to Stommel's two-box model [20]. In Tokmakian *et al.* [21], the authors discuss the use of Gaussian process emulators for nonlinear systems and exemplify their methods using Stommel's model Stommel [20]. Zickfield's model consists of four well-mixed boxes that represent the southern (box 1), northern (box 2), tropical (box 3) and deep Atlantic (box 4) as shown in Figure 1. Each box is associated to its volume V_i which is considered constant over time, temperature T_i, and salinity S_i. The temperatures T_i^*, $i = 1, 2, 3$, are the relaxation temperatures for the northern, southern and tropical boxes respectively. These temperatures represent the oceanic temperature in the absence of heat transport. In Figure 1, F_1 and F_2 are the freshwater fluxes into the tropical and northern Atlantic, and the thick black arrows represent the meridional flow.

The meridional volume transport or overturning, m, is proportional to the difference between densities in the northern and southern boxes and, therefore, is determined by the southern and northern boxes' temperatures and salinities. At a given time, the overtuning m can be calculated as follows

$$m = k\left[\beta(S_2 - S_1) - \alpha(T_2 - T_1)\right] \tag{3}$$

where k is an empirical flow constant, β is a haline expansion coefficient and α is a thermal expansion coefficient. This four-box system is modelled using the set of differential equations in (4) below

$$\begin{aligned}
\frac{dT_1}{dt} &= \frac{m}{V_1}(T_4 - T_1) + \lambda_1(T_1^* - T_1) \\
\frac{dT_2}{dt} &= \frac{m}{V_2}(T_3 - T_2) + \lambda_2(T_2^* - T_2) \\
\frac{dT_3}{dt} &= \frac{m}{V_3}(T_1 - T_3) + \lambda_3(T_3^* - T_3) \\
\frac{dT_4}{dt} &= \frac{m}{V_4}(T_2 - T_4) \\
\frac{dS_1}{dt} &= \frac{m}{V_1}(S_4 - S_1) + \frac{S_0 F_1}{V_1} \\
\frac{dS_2}{dt} &= \frac{m}{V_2}(S_3 - S_2) + \frac{S_0 F_2}{V_2} \\
\frac{dS_3}{dt} &= \frac{m}{V_3}(S_1 - S_3) + \frac{S_0(F_1 - F_2)}{V_3} \\
\frac{dS_4}{dt} &= \frac{m}{V_4}(S_2 - S_4).
\end{aligned} \tag{4}$$

In (4), S_0 is the reference salinity for the system and λ_i, $i = 1, 2, 3$, are the coupling constants for boxes 1 to 3. Each coupling constant is a function of the box's depth z_i, the specific heat capacity of sea water c, seawater's density ρ_0, and the thermal coupling Γ as follows

$$\lambda_i = \frac{\Gamma}{c\rho_0 z_i}. \tag{5}$$

The thermal coupling constant is a combination of the radiative relaxation and atmospheric heat diffusion constants as described in Zickfield *et al.* [24].

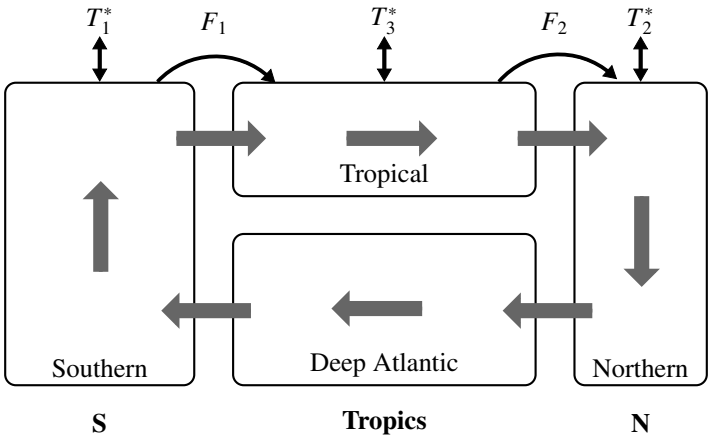

Figure 1. Four-box model of the Atlantic thermohaline circulation as proposed by Zickfield *et al.* [24]

The system is driven by a large number of variables and constants, but we are here interested in the overturning m. In order to simplify the problem, we will treat m as a function of time t, the freshwater flux from the southern box into the tropical box F_1, and the thermal coupling Γ. All the other parameters and constants are left fixed at similar values to the ones listed in [24]. Referring back to Section 2, in summary, the physical system we are describing is the Atlantic thermohaline circulation, which is modelled by a system of differential equations as in (4) and we

are only interested in one output of the simulator, m, the meridional volume transport, that we treat as a function of time, F_1, and Γ. This example is computationally cheap but its simplicity allows us to illustrate important Bayesian techniques that can be applied whenever we want to learn more about the uncertainties around the physical system to be modelled, or simply need to explore the simulator itself further without wasting resources. Dimensionality, sparsity and non-linearity are issues that often arise in systems modelled by computer simulators; the tools described in the following sections can be used to improve our understanding of the simulator and therefore the physical system. Next, we present the basic concepts of emulation.

4. Representing beliefs about f using emulators

We would like to use observations, $\mathbf{z}$, according to relations (1) and (2), in order to learn about $\mathbf{x}^*$. For most problems of realistic size and complexity, this raises difficulties as the evaluation of $f(\mathbf{x})$ for any $\mathbf{x}$ may be very expensive in time and computer resource. To address this problem, we often proceed by constructing an emulator for the simulator [4, 3, 15]; see [2, 13, 5, 23, 22] for examples of emulators used in different applications.

An emulator is a probabilistic belief specification, expressing uncertainty judgements for a deterministic function. A common choice for the form of the emulator, for component i of f, is

$$f_i^*(\mathbf{x}) = \sum_j \beta_{ij}\, g_{ij}(\mathbf{x}) + u_i(\mathbf{x}) \tag{6}$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of $\mathbf{x}$, and $u(\mathbf{x})$ is a stationary stochastic process. A common choice is to suppose that $u(\mathbf{x})$ is a Gaussian process, so that, for each $\mathbf{x}$, $u_i(\mathbf{x})$ is normal with constant variance and $\mathrm{Corr}(u_i(\mathbf{x}), u_i(\mathbf{x}'))$ is a function of $\|\mathbf{x} - \mathbf{x}'\|$.

We fit the emulator, given the collection of model evaluations $\mathbf{F}$, using appropriate statistical tools, such as generalised least squares, maximum likelihood, Bayes, combined with detailed expert judgement, where available. The form of this emulator will be illustrated in detail, in the context of Zickfield's model, in Section 6. The methodology that has been developed for emulator construction is based on careful experimental design to choose which evaluations of the model to make, and detailed diagnostics, to check emulator validity. If the simulator is very expensive to evaluate then we often start by making many evaluations of a simpler approximate version of the simulator from which we may develop an informative prior for the emulator of the full simulator [6].

5. Bayes linear analysis and history matching

A full Bayesian analysis of a complex physical system, based on a computer simulator, requires a complete specification of (i) a prior probability distribution for inputs $\mathbf{x}^*$, (ii) a probabilistic emulator for the simulator f, (iii) a probabilistic discrepancy measure relating $f(\mathbf{x}^*)$ to the system $\mathbf{y}$, (iv) a likelihood function relating historical data $\mathbf{z}$ to $\mathbf{y}$.

This detailed description provides a formal framework to synthesise expert elicitation, historical data and simulator evaluations, and thus to carry out a full uncertainty analysis for the physical system.

For relatively small systems, this approach is practical and successful [14]. For large problems, however, it is difficult to provide a meaningful prior specification expressing all of the relevant uncertainties in the problem. Also the resulting analysis usually is extremely computationally intensive and potentially highly sensitive to the precise form of the prior specification.

In such circumstances, there is an alternative method for specification and analysis of uncertainties, which is similar in spirit to full Bayes but which is based directly on expectation as a primitive, involving prior specification of means, variances and covariances alone. Working directly with expectations greatly simplifies both the uncertainty specification and the analysis. There are sound theoretical foundations for the expectation based approach.

The statistical approach in which expectation is treated as primitive is termed Bayes linear analysis; for an overview of the notion of treating expectation as primitive, see [7], for a summary overview of the Bayes linear approach, see [8] and for a full account of the Bayes linear approach, see [12]. The approach is based around the following updating equations which give the adjusted mean and variance of vector y, given observation of vector z, based on prior means, variances and covariances for each quantity which are specified directly by whatever methods

are appropriate for the problem at hand (for example, in the problems that we are considering, through the construction of an emulator).

$$\mathrm{E}_{\mathbf{z}}(\mathbf{y}) = \mathrm{E}(\mathbf{y}) + \mathrm{Cov}(\mathbf{y}, \mathbf{z})\mathrm{Var}(\mathbf{z})^{-1}(\mathbf{z} - \mathrm{E}(\mathbf{z})), \tag{7}$$

$$\mathrm{Var}_{\mathbf{z}}(\mathbf{y}) = \mathrm{Var}(\mathbf{y}) - \mathrm{Cov}(\mathbf{y}, \mathbf{z})\mathrm{Var}(\mathbf{z})^{-1}\mathrm{Cov}(\mathbf{z}, \mathbf{y}) \tag{8}$$

The examples that we will describe use Bayes linear methods, where appropriate. In particular, as an example of the application of the Bayes linear approach, we now describe a method, based on second order specifications, for identifying appropriate classes of model inputs which respect the constraints of historical data. This problem is related to, but distinct from, the problem of model calibration.

Model calibration aims to find the posterior distribution for the input value $\mathbf{x}^*$, given the data $\mathbf{z}$. We may have reservations about this approach however. Firstly, we may not believe in a unique value of best input value for the simulator. Indeed, we may be unsure as to whether any values of input parameters would enable the simulator to match the physical system. Secondly, such probabilistic calibration analysis may be difficult and non-robust [9].

Therefore, we may prefer to use a procedure termed 'history matching' [22], namely finding the collection, $C(\mathbf{z})$, of all input values $\mathbf{x}$ for which the match of the model outputs $f_h(\mathbf{x})$ to observed data, $\mathbf{z}$, is judged to be acceptably small, taking into account all of the uncertainty. If $C(\mathbf{z})$ is non-empty, then an analysis of its form may reveal the constraints on the parameter space imposed by the data. Further, the simulator evaluations $f(\mathbf{x}) : \mathbf{x} \in C(\mathbf{z})$ for future time points, reveal the futures consistent with the model physics and the historical data. If the data is informative for the parameter space, then $C(\mathbf{z})$ will typically form a tiny percentage of the original parameter space, so that, even if we do wish to calibrate the simulator, history matching is usually an important first step.

We search for $C(\mathbf{z})$ by seeking to remove from the parameter space input values which we consider unlikely to be members of that set. We do this by use of an 'implausibility measure' $I(\mathbf{x})$ which uses a metric based on the number of standard deviations between $\mathbf{z}$ and $f_h(\mathbf{x})$, based on the formulation (1), (2). For example, if we are matching a single output, then we might choose

$$I(\mathbf{x}) = \frac{(\mathbf{z} - \mathrm{E}(f_h(\mathbf{x})))^2}{\mathrm{Var}(\mathbf{z} - \mathrm{E}(f_h(\mathbf{x})))} \tag{9}$$

where, from (1), (2)

$$\mathrm{Var}(\mathbf{z} - \mathrm{E}(f_h(\mathbf{x}))) = \mathrm{Var}(\mathbf{e}) + \mathrm{Var}(\epsilon) + \mathrm{Var}(f_h(\mathbf{x})). \tag{10}$$

The implausibility calculation can be performed univariately, or by multivariate calculation over sub-vectors. Typically, we do not attempt to match the whole history, $\mathbf{z}$, at this stage, but rather a subvector $z_{(1)}$ of those outputs which can each be emulated effectively, and which individually provide different constraints on the parameter space. The implausibilities are then combined, such as by using $I_M(\mathbf{x}) = \max_i I_{(i)}(\mathbf{x})$, and can then be used to identify regions of $\mathbf{x}$ with large $I_M(\mathbf{x})$ as implausible, i.e. unlikely to be good values for $\mathbf{x}^*$. If we do not wish to make detailed probabilistic assumptions, for example, if we are carrying out a Bayes linear analysis based only on means, variances and covariances, then we may employ the three sigma rule (which says that for any continuous unimodal distribution at least 95% of the probability is within three sigma of the mean) as a guide to setting cut-offs for the implausibility measure Pukelsheim [16].

The result of this analysis is to identify a region $C^*_{(1)}(\mathbf{z}_{(1)})$ of non-implausible input values, based on history matching using $\mathbf{z}_{(1)}$. With this information, we may refocus our analysis on $C^*_{(1)}(\mathbf{z}_{(1)})$ by (i) making more simulator runs within this sub-region and (ii) refitting our emulators over this sub-region and repeating the analysis. This process is a form of iterative global search. At this stage, we often find that outputs which could not be well emulated in the original parameter space become smoother and more predictable in the reduced space, so that we are able to expand the subvector $\mathbf{z}_{(1)}$ to a larger subvector $\mathbf{z}_{(2)}$. With this process, we reduce $C^*_{(1)}(\mathbf{z}_{(1)})$ to a sub-region $C^*_{(2)}(\mathbf{z}_{(2)})$. We continue to refocus in waves, in this fashion, until we have sufficiently reduced the volume of the non-implausible region for the purpose of the study.

6. Emulating the four-box model

In Section 3, we described a four-box model for the Atlantic thermohaline circulation. The model is represented by a set of differential equations in (4) and we are interested in investigating the meridional overturning m in (3) as a function of time t, freshwater forcing F_1, and thermal coupling Γ. We fix the model's constants and start points using the recommended values for present-day climate in Table 1 of Zickfield *et al.* [24]; likewise, we choose appropriate ranges for F_1 and Γ as $(-0.2, 0.2)$Sv and $(10, 75)$ Wm^{-2}, respectively, in accordance with the relevant discussion in that paper. Figure 2 shows the response of the overturning for different time points. Note how the surface starts as a smooth polynomial and then slowly folds into a discontinuity.

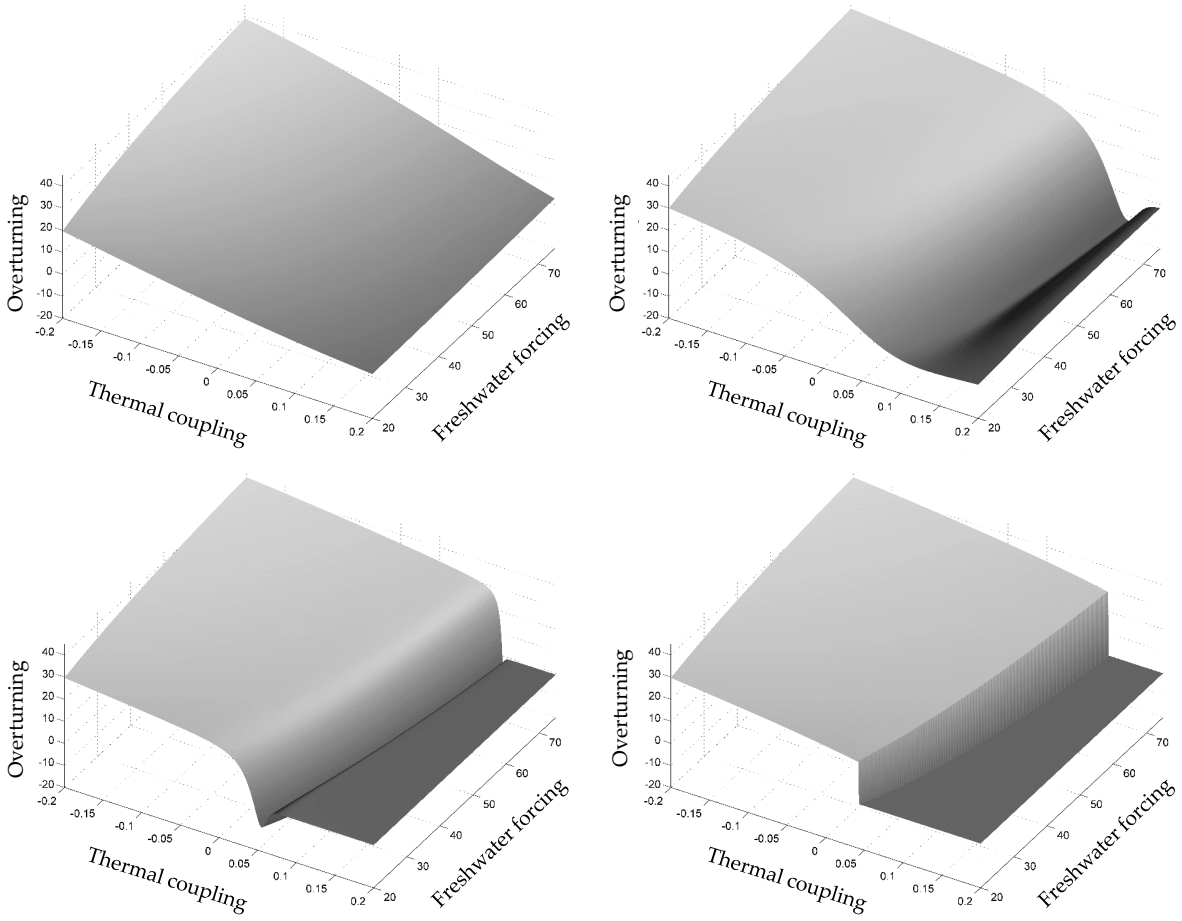

Figure 2. Surfaces representing the overturning m as a function of freshwater forcing F_1 and thermal coupling Γ at times $t = 50$ (top left), $t = 100$ (top right), $t = 300$ (bottom left), and $t = 1000$ (bottom right).

To illustrate the process of emulation, suppose that the simulator is expensive to evaluate. To avoid wasting resources, the simulator runs have to be chosen through a carefully designed experiment. Once the input space is specified, as above, a sampling method and frame should be specified to achieve good input space coverage. A popular method is Latin hypercube sampling which is a form of stratified sampling that imposes an even distribution of points over the marginals of the input space [19]. For this exercise, we start by generating 25 samples in $(-0.2, 0.2)\times(10, 75)$, as shown in Figure 7(a), using a Latin hypercube design with a minimum correlation criterion which minimizes the cross-correlation across the columns of the design. These samples are plotted in Figure 3.

We are interested in building emulators to represent the overturning surface at a given time t, i.e. $m(t) = f_t(F_1, \Gamma)$.

For fixed t, we write

$$f_t^*(\mathbf{x}) = \sum_j \beta_{tj}\, g_{tj}(\mathbf{x}) + u_t(\mathbf{x}), \quad \text{where } \mathbf{x} = (F_1, \Gamma). \tag{11}$$

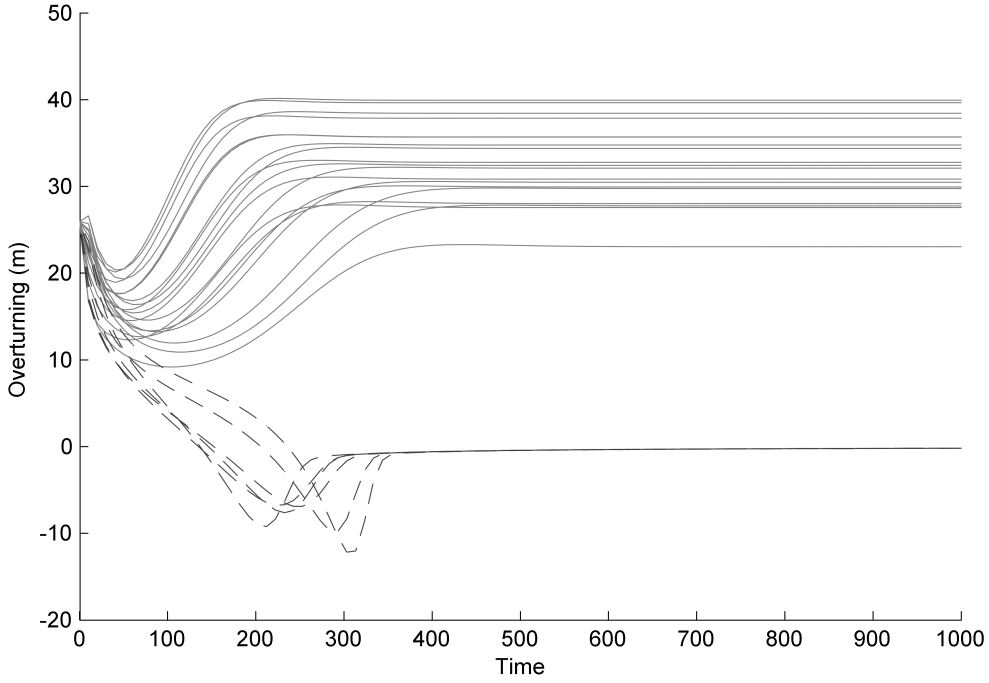

Figure 3. Simulator runs for different parameter values represented as time-series of overturning (m) as a function of time t obtained by simulating the four-box model on each of the 25 samples. The solid gray curves show the cases where $m > 0$ at equilibrium, and the dashed dark gray curves, when $m \leq 0$ at equilibrium.

We choose as the kernel of $f_t^*(\mathbf{x})$, $\sum_j \beta_{tj}\, g_{tj}(\mathbf{x})$, a second-degree polynomial in F_1 and Γ so that

$$\sum_j \beta_{tj}\, g_{tj}(\mathbf{x}) = \beta_{t0} + \beta_{t1} F_1 + \beta_{t2}\Gamma + \beta_{t3} F_1\Gamma + \beta_{t4} F_1^2 + \beta_{t5}\Gamma^2 \tag{12}$$

and choose $u_t(\mathbf{x})$ as a second-order stationary process with zero-mean and squared-exponential covariance function, i.e.

$$Cov(u_i(\mathbf{x}), u_i(\mathbf{x}')) = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\theta^2}\right).$$

We will not discuss choices of alternative functional forms for the covariance function (see Rasmussen and Williams [17] for further discussion), except to note that the chosen functional should represent one's beliefs on stationarity, smoothness, and symmetry.

We fit the emulator described above using 25 samples at time $t = 40$. The kernel must be chosen so that the main response is well represented without over-fitting. Here we fitted the response with polynomial terms. We fitted emulators with linear (adjusted $R^2 = 0.9876$, RMSE = 0.3566), quadratic (adjusted $R^2 = 0.9996$, RMSE = 0.07258), and third-order kernels (adjusted $R^2 = 0.9998$, RMSE = 0.01455), and we noticed that there was very little gain with the addition of third-order elements. The most relevant term is the interaction between Γ and F_1 which cannot be fully absorbed by the covariance function alone; the second and third-order emulators with interaction terms were easily validated unlike the ones without mixed terms. The validation was performed using diagnostic methods discussed later in this Section. After choosing a suitable kernel, we estimate $B = \{\beta_{ij}\}$ using least squares. From the regression

residuals, we construct the posterior distribution of σ^2 and θ. The residuals are calculated using the least-squares fitted $\hat{B}$. We then calculate the maximum a posteriori estimates for σ^2 and θ. For this calculation, we assumed that the underlying process is a Gaussian process in order to construct the likelihood. Alternatively, B, σ^2 and θ could be simultaneously assessed by constructing a joint posterior distribution; however, specification of the covariance structure between all of the parameters can be challenging. In general, assessing the covariance parameters fore the residual process is a challenging problem. As, in our case, the residual variance is small, we use the simple heuristic estimate corresponding to the maximum a posteriori estimates for σ^2 and θ, assuming a Gaussian likelihood. However, we shall not use the Gaussian assumptions in our subsequent analysis, and so we validate our emulator under the weaker second order assumptions by application of the three-sigma rule (see Pukelsheim [16]).

Our simulator is computationally cheap so we additionally simulated the whole surface at $t = 40$ to draw a comparison between the simulated surface and the posterior mean of the emulator. The simulated surface is displayed in Figure 4, with the difference between simulated and posterior mean surface and their corresponding standard errors. We can see that the difference between the simulated surface and the expected value of the emulator is small suggesting a good approximation. Here we included the complete surface to provide a complete demonstration to the reader that, in this case, the emulator really does approximate the model well. When dealing with a real problem, we would not usually be able to simulate the entire surface for comparison; we can, however, make some extra function evaluations and run diagnostic tests. Calculating individual standardised errors is a simple test that consists of computing the ratio of the difference between a simulated value and the emulator's expected value and the emulator's corresponding a-posteriori standard deviation, i.e.

$$se = \frac{f(\mathbf{x}) - E(f^*(\mathbf{x}))}{sd(f^*(\mathbf{x}))}. \tag{13}$$

Usually extreme absolute standard errors (much greater than 3, see Pukelsheim [16]) or clusters of extreme values indicate problems with the emulator. In Figure 4, the rightmost graph shows the standard error surface for our emulator. Here the standard errors are well within the interval $(-3, 3)$, for all x. As a validation exercise, we selected 10 new samples placed on the neighbourhood of the original points and 10 other samples placed closer to the boundaries of the input space. The calculated standard errors for these samples were within the expected interval $(-3, 3)$; the samples closer to the original points had small standard errors while the points closer to the boundaries showed slightly larger discrepancies. The larger the number of validation tests that the emulator passes, the more confident that we will be that the emulator does a satisfactory job in representing our uncertainty about the function over the parameter range of interest. See Bastos and O'Hagan [1] for a more detailed treatment of diagnostics for emulators, taking into account the correlation between the residuals.

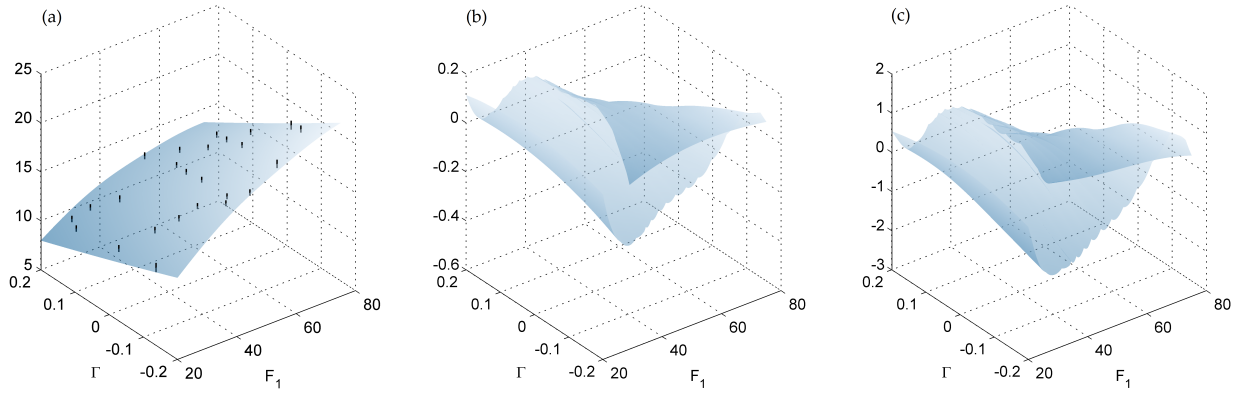

Figure 4. (a) Simulated surface for overturning m at $t = 40$, (b) difference between simulated values and the emulator's posterior mean, and (c) corresponding standardised errors calculated as in (13).

Emulation is used for building local and global approximations for a simulator. We have illustrated that, with a fairly small dataset of simulator evaluations, we can approximate the true simulator surface. Unlike some popular

methods for interpolation, like splines, an emulator is a probabilistic approximation of the simulator and therefore carries an uncertainty representation which we can use to assess the quality of this approximation for any point of the input space at a very low cost. Emulators are also often used for model calibration; in our example, we could be looking for the 'best' model input given a set of observed values of m, or we might want to know if the parameter space can be classified into distinct subregions according to given criteria. One can assess this through a full calibration analysis [14], which often is very expensive computationally, or via history matching, as we now describe.

7. History matching the four-box model

We now demonstrate how, in the example, history matching can be used to eliminate implausible regions in the input space. We select one point in the input space $(F_1, \Gamma) = (-0.1, 45)$, which we assume to be the true value for (F_1, Γ), and simulate the four-box model for time $t \in [0, 1000]$. We then add noise to the simulated time-series in the form of white noise to represent the measurement error in (2) with zero mean and variance $0.01E(m(t, \mathbf{x}))$, and correlated noise to account for the structural discrepancy in (1) with zero mean, and covariance $0.03 \min(E(m(t_1, \mathbf{x})), E(m(t_2, \mathbf{x})))$. The result is plotted in Figure 5, the black solid line shows the simulated path and the gray line shows the simulated path with added noise. We observe values of m on the 'noisy' curve at different times, compute the implausibility score $I(\mathbf{x})$ using the emulator in (12), for all $\mathbf{x}$ as described in 9. This emulator is fitted using the 25 paths associated to the input points displayed in Figure 7(a). We use the cut-off of 3 (i.e. 3 standard deviations) meaning that, for a given $\mathbf{x}$, if $I(\mathbf{x}) > 3$, then $\mathbf{x}$ is discarded as an implausible input.

In order to calculate implausibilities, we need to specify observational error and structural discrepancy. In the first case, we assume that structural discrepancy is nonexistent and that the overall error (observational error and structural discrepancy) is normally distributed with zero-mean and variance $0.05E(m(t, x))$.

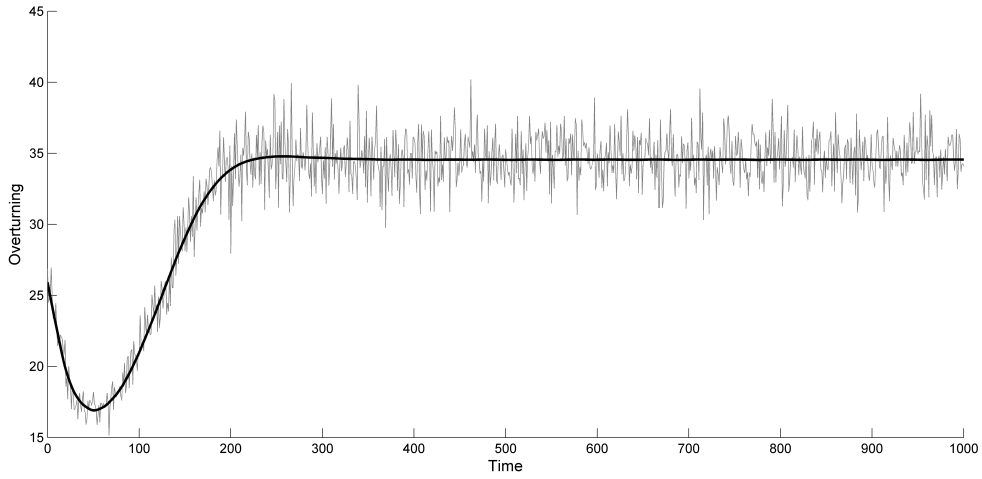

Figure 5. Sample path for history matching exercise. Solid black line: simulated sample path, gray line: sample path with correlated noise.

On the sample path in Figure 5, we observe m at $t = 40$, and compute the implausibility $I(\mathbf{x})$ on a grid of the input space. In Figure 6, plot (a) shows the implausibility map with a cut-off of 3 for the case where we haven't taken observational error into account; the shaded area represents the non-implausible region in the input space, and the true input value is indicated by the dashed lines. Note that the true value is outside the non-implausible region. Plot (b) shows the implausibility map, also with a cut-off of 3, with observational error accounted for. The resulting non-implausible region is shown in Figure 6 (b); the new region is considerably larger and contains the true value. In both cases, the non-implausible regions are concentrated around a curve instead of a point. Indeed, for this problem, there are an infinite number of points in the input space laying on a level curve that map back to the observed value for m at $t = 40$. We add another wave of 25 samples to the non-implausible region in (b) as shown in Figure 7, refit

the emulator in (12) and recalculate the implausibility map obtaining the region in plot (c). In plot 6(c), we see that the non-implausible region is thinner than the one obtained after the first wave in (b) and the true value is closer to the centre. In plot (b), the input space is reduced to 18% of its original size, and, with the second wave, in plot (c), the non-implausible regions is reduced by 28%, i.e. an extra 5% of the original input space.

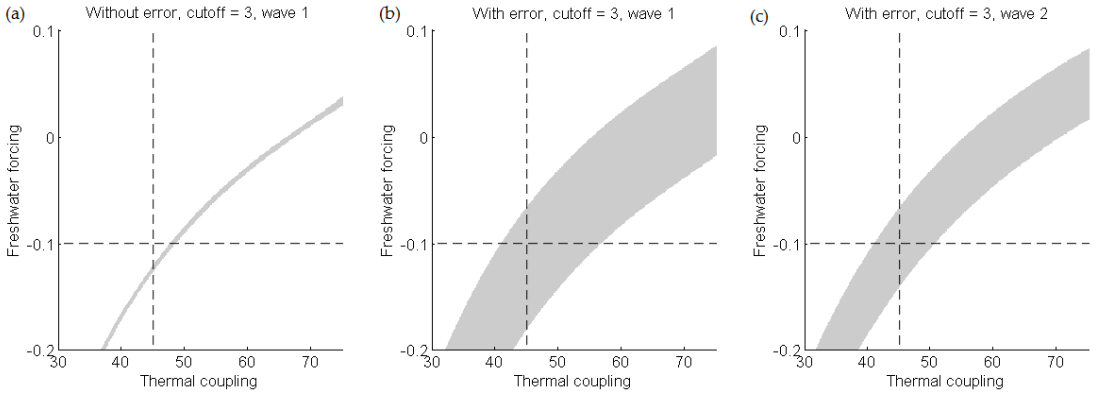

Figure 6. Implausibility maps with cut-off 3 for the sample path in Figure 5 at $t = 40$ using the 25 samples plotted in Figure 3 to build the emulator (a) without taking observational error into account, and (b) accounting for observational error, reducing the input space by 82%. We then added 25 more samples to the non-implausible region in (b) in a second wave and produced the region in (c) which corresponds to 72% of the non-implausible region in (b) and 13% of the original input space.

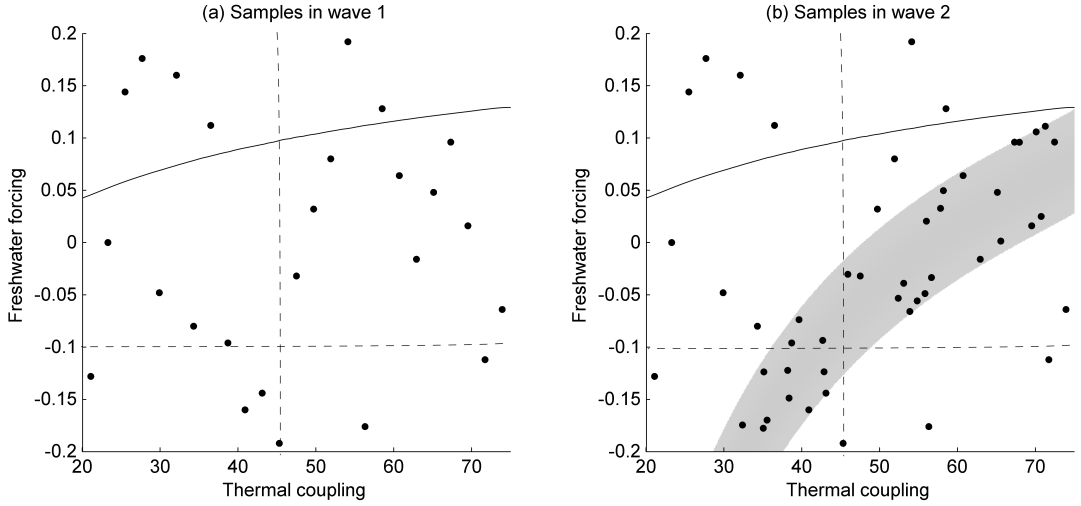

Figure 7. (a) Samples in wave 1 used to fit the emulator at $t = 40$. (b) Samples in wave 1 and 2 randomly taken from the non-implausible area in Figure 6(b). The black curve line represents the boundary between the collapse and non-collapse regions, i.e. $m \leq 0$ above this boundary and $m \geq 0$ otherwise. The "true" value used for history matching lays on the intersection of the two dashed lines.

Now we improve our assessment of the overturning by further observing m at a second time point. We observe m at $t = 90$ and we build an emulator for the difference $d(90, \mathbf{x}) = m(90, \mathbf{x}) - m(40, \mathbf{x})$, using the original 25 observations. We then construct the corresponding implausibility measure accounting for observational error, assumed to have zero-mean and variance $0.01E(m(\mathbf{x}, t))$, and structural discrepancy, with zero mean, and covariance $0.03 \min(E(m(t_1, \mathbf{x})), E(m(t_2, \mathbf{x})))$. We look at the maximum implausibility $I_M(\mathbf{x})$ for $m(40, \mathbf{x})$ and $d(90, \mathbf{x})$, and add 25 more samples to the non-implausible regions.

In Figure 8, we show the non-implausible region when observational error is taken into account (left), and the non-implausible region for the case with both observational error and structural discrepancy (right). Both non-implausible

regions have roughly the same area. However, the second shows that the shape of the first region was mostly driven by unresolved interactions between variables across time, i.e. the variance for each variable is well-represented but not their covariance structure. By accounting for structural discrepancy, the true value now sits closer to the centre of the non-implausible area instead of the edge, and the marginal implausibilities cover a smaller region of the input space reducing uncertainty on the value of thermal coupling.

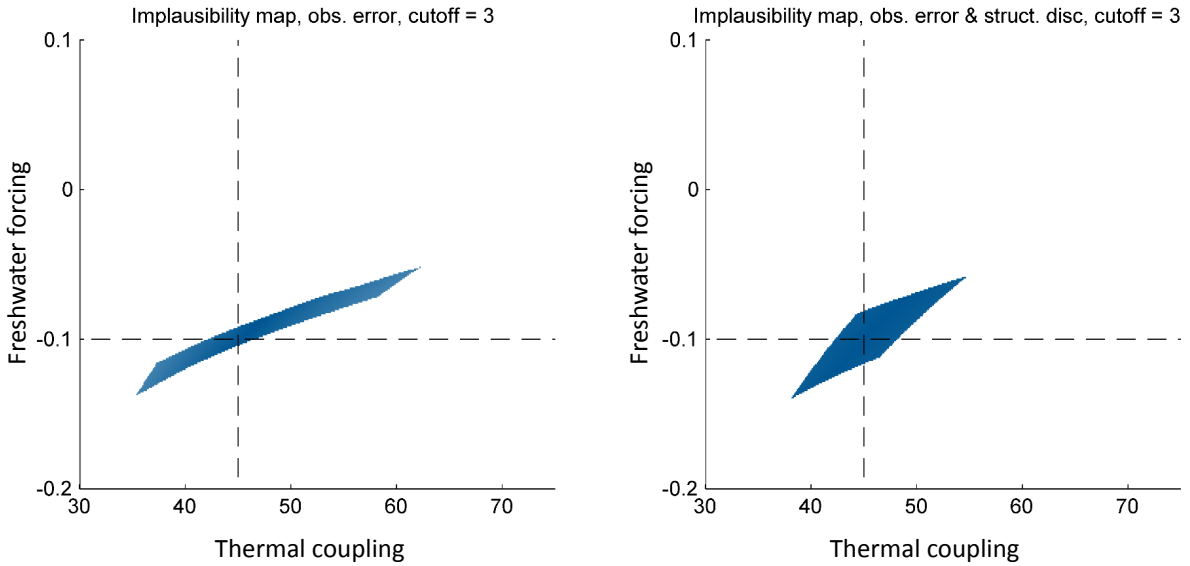

Figure 8. Implausibility maps with cut-off 3 for the sample path in Figure 5 for $m(40, \mathbf{x})$ and $m(90, \mathbf{x}) - m(40, \mathbf{x})$ accounting for observational error only (left), and observational error and structural discrepancy (right) after two waves.

8. History matching with tipping points

The methods that we have described involve emulation of the computer simulator, treated as a smooth function of the inputs. From now on, we suppose that the simulator output is a time series $f(\mathbf{x}, t)$, with more than one form of limiting behaviour (e.g. collapse or non-collapse of the physical system), so that the input space divides into regions R_i such that simulator output is smooth within each region, but discontinuous across boundaries. We suppose that we can discover which region any input $\mathbf{x}$ belongs to by running the simulator to equilibrium for that value of $\mathbf{x}$. Typically, this will be an expensive operation, so that we will only be able to make such an assessment for a limited sample of input values. We shall describe our procedure for the problem with two regions (which we may think of as system collapse or system non collapse), but the argument generalises to many regions similarly.

We modify our procedure as follows. We make a sample of input values from each region R_i. We cannot do this directly, as we do not know which region each $\mathbf{x}$ value belongs to, and thus we must evaluate a carefully selected collection of $\mathbf{x}$ values and divide them into regions by observation of their equilibrium behaviour.

This is not problematic if the function is relatively cheap to evaluate and no region is small as compared to the other regions. When these conditions fail, we will need to sample carefully, guided by physical intuitions as to likely parameter configurations leading to the different limiting states, possibly guided by a faster, approximate model with similar limiting behaviour. This raises interesting methodological issues, which we shall explore in detail elsewhere. Here, our purpose is simply to show how the ideas of emulation, model discrepancy and history matching may be carried over directly to problems exhibiting tipping point behaviour.

Thus, we suppose that we may obtain two samples of model evaluations, one from each region. We choose an initial subvector $\mathbf{z}_{(1)}$ for our first stage history match. From the two samples, we construct two emulators $f^*_{[1]}, f^*_{[2]}$ for the corresponding outputs, one for each region of input space. We therefore have two implausibility scores, $I_{[1]}(\mathbf{x}), I_{[2]}(\mathbf{x})$ for $\mathbf{x}$, one for each emulator.

When evaluating the implausibility measures, we must consider carefully the specification of structural discrepancy variance in the relation (1). We may consider that the ability of the model to capture real system behaviour differs between regions of the parameter space; for example, we might consider that the model represents reality well in the non-collapse region, but is less reliable in tracking system collapse, if it should occur. We will therefore often choose different values of discrepancy variance for $mathbf\epsilon$, in (10), depending whether we are evaluating $I_{[1]}(\mathbf{x})$ or $I_{[2]}(\mathbf{x})$.

Given our two implausibility measures, we may construct two regions of non-implausible input values $C^*_{[i](1)}(\mathbf{z}_{(1)}), i = 1, 2$. For a general input, $\mathbf{x}$, we do not know which region it belongs to. Therefore, in general, we can only remove an input $\mathbf{x}$ from the input space if it is not a member of either subspace $C^*_{[i](1)}(\mathbf{z}_{(1)}), i = 1, 2$. Similarly, we may only reject all values of $\mathbf{x} \in R_i$ and therefore reject the form of behaviour (collapse or not) corresponding to that region, if the corresponding subspace $C^*_{[i](1)}(\mathbf{z}_{(1)})$ is empty.

Just as for the standard history match, we may continue by refocussing in waves within each of the two sub-regions, producing a decreasing sequence of subsets of input space $C^*_{[i](j)}(\mathbf{z}_{(j)}), i = 1, 2, j = 1, 2, ...$ Again, we cannot re-sample these subsets directly. Instead we must devise a general sampling strategy for $\mathbf{x}$ at each stage, and allocate the observed function evaluations to the appropriate regions given the corresponding equilibrium behaviour. This allows us to shrink the pair of regions progressively until we reach a stopping criterion appropriate to the aims of the study.

9. Example: History matching the four-box model with tipping points

In our example, we are interested in the behaviour of the system at equilibrium. For present-day climate, this four-box model presents a bifurcation that can lead to vigorous overturning, $m_{eq} > 0$, the preferred scenario for future climate, or the collapse of the Meridional overturning circulation, $m_eq \leq 0$, the scenario where the water masses in the southern box sink and the northern box up-wells. These two scenarios map back to two regions in the input space R_1, the non-collapse zone where $m > 0$, and R_2, the collapse zone where $m \leq 0$. The fourth surface in Figure 2 (bottom right) clearly show the discontinuity separating regions R_1 and R_2. Each region is well defined and the function is well-behaved within each region. Because of the discontinuity across the boundary between regions R_1 and R_2, we do not create a single emulator for the combined region. Instead, we have a sample of function evaluations known to be from R_1, which we use to build an emulator for the function over R_1 and a sample from region R_2, which we use to build an emulator for the function over R_2. Although each emulator is valid for the appropriate region, we do not know, for a general value of input parameters, which region it belongs to, unless we evaluate the function at this value. Therefore, when we apply our history matching procedure, we must consider the possibility that the value is in each of the regions, apply the appropriate emulator for that regions to this value, and draw the appropriate conclusions given the pair of inplausibility evaluations.

Using the sample path in Figure 5 as the true path, we now suppose that we observe values of m at different times, compute the implausibility scores $I_{[1]}(\mathbf{x})$ and $I_{[2]}(\mathbf{x})$ for all $\mathbf{x}$ as in (9). Analogously to the example in Section 7, we use the cut-off of 3 (i.e. 3 standard deviations) meaning that, for a given $\mathbf{x}$, if $I_{[i]}(\mathbf{x}) > 3$, then $\mathbf{x}$ is discarded as an implausible input for the corresponding emulator.

We start by building two emulators at $t = 5$, one for each region, using 50 samples selected using a Latin hypercube design. We run the simulator for each input for $t \in [0, 1000]$ and classify as being in region R_1 or R_2. There are 28 samples in R_1 and 22 in R_2. The two emulators have the same kernel and covariance function as in Section 6 and we fit each with samples in their corresponding regions. Observing m at $t = 5$, gives us the implausibility maps for $m(5, \mathbf{x})$ in Regions 1 and 2 shown in Figure plots 9 (a.1) and (a.2). The result for region R_1 is smaller than the one in Figure 6 as we would expect mainly because the samples used to fit the emulator in R_2 are quite far from the true value. Similar to the example in Section 7, we assume that the overall error is normally distributed with zero-mean and variance $0.05m(t, \mathbf{x})$ for both regions.

Now that the input space has been reduced significantly, we take another sample of size 25 in the union of the non-implausible areas in 9 (a.1) and (a.2). We refit the emulators for $m(5, \mathbf{x})$ and build the emulator for $m(15, \mathbf{x}) - m(5, \mathbf{x})$ in each region. We compute $I_{M[1]}(\mathbf{x})$ and $I_{M[2]}(\mathbf{x})$ for all $\mathbf{x}$, where $I_{M[i]}(\mathbf{x})$ is the maximum implausibility for $m(5, \mathbf{x})$ and $m(15, \mathbf{x}) - m(5, \mathbf{x})$; the resulting non-implausible regions are displayed in 9 (b.1) and (b.2). In this case, as we have information about 2 points in the curve, we separate the overall error into observational error and structural discrepancy; for the first we assume that the observational errors are independent and normally distributed with zero-

mean and variance $0.01E(m(t, \mathbf{x}))$, while the structural discrepancy holds a correlated structure with zero-mean, and covariance $0.03 \min(E(m(t_1, \mathbf{x})), E(m(t_2, \mathbf{x})))$ at R_1 and $0.01 \min(E(m(t_1, \mathbf{x})), E(m(t_2, \mathbf{x})))$ at R_2.

We continue this analysis by adding another wave of 25 randomly selected samples to the combined non-implausible area. Most of the non-implausible area falls within region 1 so it is more likely that randomly selected samples will belong to region 1, here 7 of the new samples fall in region 2 and 18 fall in region 1. We refit the emulators for $m(5, \mathbf{x})$ and $m(15, \mathbf{x}) - m(5, \mathbf{x})$, and we fit the emulator for $m(25, \mathbf{x}) - m(15, \mathbf{x})$ with samples of each region. The resulting regions can be seen in 9 (c.1) and (c.2). Finally, we include one more wave of 25 samples, refit the first three emulators in each region, and include emulators for $m(35, \mathbf{x}) - m(25, \mathbf{x})$ resulting in the non-implausible regions in 9(d.1) and (d.2). All samples in this final wave belong to region 1 as the non-implausible area is fully contained within this region. After the final fit, most of the points in region 2 are labeled as implausible; in fact, with one more wave, no points can be labeled as non-implausible and we are able to classify the observed path as belonging to region 1 with a very high probability.

As illustrated, by repeating the exercise of observing more points in each sample path and adding more waves, often we can eventually rule a region implausible and label the path as being in the collapse or non-collapse regions for any path. In order to illustrate this procedure further, we create a grid on the input space and take 500 simulated sample paths with added white noise to represent observational error and Brownian noise as structural discrepancy with the same variance and covariance structures describe above. For each path, we observe m at t, reduce the input space using the implausibility measures for each emulator, observe m again at $t + 1$, create a joint emulator and reduce the space further. If t is divisible by 5, we add a new wave with 10 new samples within the joint non-implausible region for each sample path. If a region (R_1 or R_2) is deemed implausible, then we stop observing m and record the time t when we were able to identify if $m_{eq} > 0$ or $m_{eq} \leq 0$. The results are displayed in Figure 10. As we would expect, it is much harder to eliminate a region near the boundary; in fact, at some points, the region can only be ruled out near the time the system would have reached the bifurcation point. This raises interesting points related to decision making, if we were in the collapse region and near the boundary, it is unlikely we would be able to establish that we were heading to collapse until it became inevitable; under these conditions, it would be reasonable to take preventative action. Similarly, if we could establish that we were well within one region or another, it is still possible that unforeseen actions could lead to variations in freshwater forcing and thermal coupling beyond the scope of any model; these undetected uncertainties could ultimately result in a different equilibrium state.

10. Conclusion

We have shown how emulation and history matching can be used to simplify the analysis of a complex numerical simulator, illustrating how to link a simulator to the real world by incorporating observational errors and structural discrepancies; emulation can be used to create global and local approximations for complex systems, and history matching may be used to explore and reduce the input space. There are many further aspects of uncertainty that one could explore using these methods. For example, in the particular problem of the four-box model, one could investigate the uncertainty around the simulator's initial conditions, the functional choice in the system of differential equations, the underlying stochasticity that is not represented by a deterministic simulator as the one proposed, among others [11]. We also presented a small example on classification for tipping points using history matching which can be easily modified to include more than two regions. Incorporating aspects of decision making to models that involve classification requires a higher level synthesis of the ingredients that we presented. Such synthesis is beyond the scope of this paper, and will be addressed in future work.

Another discussion to be further developed is the reification of the four-box model. A conventional Bayesian analysis assumes that there is a unique, true but unknown, value of the parameter. As the value is unknown, a prior probability distribution is specified over the possible values of the parameter, which is converted, given the data, to a posterior distribution, via Bayes theorem. Therefore, while there is uncertainty as to what the value of the parameter is, the formal rationale for the Bayesian analysis is that one of the values of the parameter is indeed true, and has generated the data that we have observed. However, when the model is not exactly correct, and the data has not been generated by one of the choices of the parameter, then the status of the Bayesian analysis is far less clear, hence, for example, the difficulties in interpretation of Bayesian model averaging, in problems where the "true" model is not of one of the models being averaged.

Often, when we think deeply about the relationship between our model and reality, we are reluctant to assert that one of the parameter values is true. Rather, we may take the view that there is a class of parameter values for which the model outcomes may be informative for future system behaviour, hence our preference for history matching over calibration. Of course, this does raise deep questions about our aims in construction and analysing scientific models, which are beyond the scope of this paper to address. One approach for producing a more careful description of the relationship between system properties and system behaviour, and thus allowing us to ascribe a clearer meaning to the notion of true values for input parameters for our models, is termed reified modelling (from reify - to treat an abstract concept as if it were real). Goldstein and Rougier [10] give a general account of reified modelling, using the Zickfield four-box model, evaluated at equilibrium, to illustrate this process.

11. Acknowledgements

This work is part of the Durham Tipping Points project funded by the Leverhulme Trust.

References

[1] L. S. Bastos and A. O'Hagan, 2008, Diagnostics for Gaussian process emulators, *Technometrics*, **51**:425-438.
[2] P. G. Challenor, R. K. S. Hankin, and R. Marsh, 2006, Towards the probability of rapid climate change, in H. J. Scellnhuber *et al.* (eds.). *Avoiding Dangerous Climate Change*, Cambridge University Press, Cambridge.
[3] S. Conti and A. O'Hagan, 2010, Bayesian emulation of complex multi-output and dynamic computer models, *Journal of Statistical Planning and Inference*, **140**(3):640-651.
[4] S. Conti, J.P. Gosling, J.E. Oakley, and A. O'Hagan, 2009, Gaussian process emulation of dynamic computer codes, *Biometrika*, **96**:663-76.
[5] P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith,1997, Pressure matching for hydocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion). in C. Gastonis *et al.*, *Case Studies in Bayesian Statistics*, **III**, 37-93. Springer-Verlag.
[6] J. Cumming and M. Goldstein, 2009, Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations, *Technometrics*, **51**(4):377-388.
[7] M. Goldstein, 1986, Prevision, *Encyclopaedia of Statistical Sciences*, S. Kotz and N. L. Johnson eds., **7**:175-6, Wiley.
[8] M. Goldstein, 1999, Bayes linear analysis, *Encyclopaedia of Statistical Sciences*, S. Kotz, C. B. Read and D. L. Banks eds., **3**:29-34, Wiley.
[9] M. Goldstein and J.C. Rougier, 2006 Bayes Linear Calibrated Prediction for Complex Systems, *Journal of the American Statistical Asociation*, **101**(475):1132-1143.
[10] M. Goldstein and J.C. Rougier, 2009 Reified Bayesian Modelling and Inference for Physical Systems, *Journal of Statistical Planning and Inference*, **139**(3):1221-1239.
[11] M. Goldstein, A. Seheult, and I.R. Vernon, 2013 Assessing Model Adequacy, in *Environmental Modelling: Finding Simplicity in Complexity*, Second Edition, Wiley.
[12] M. Goldstein and D.A. Wooff, 2007 *Bayes Linear Statistics: Theory and Methods*, Chichester, John Wiley.
[13] R.G. Haylock and A. O'Hagan, 1996 On inference for outputs of computationally expensive algorithms with uncertainty on the inputs, *Bayesian Statistics 5*, Oxford University Press, Oxford.
[14] M. Kennedy, and A. O'Hagan, 2001 Bayesian calibration of computer models (withdiscussion), *Journal of the Royal Statistical Society:Series B*, **63**:425-464.
[15] A. O'Hagan, 2006 Bayesian analysis of computer code outputs: a tutorial, *Reliability Engineering and System Safety*, **91**:1290-1300.
[16] F. Pukelsheim, 1994 The Three Sigma Rule, *American Statistician*, **48**(2):88-91.
[17] C. E. Rasmussen and C. Williams, 2006 *Gaussian Processes for Machine Learning*, MIT Press.
[18] J.C. Rougier, 2007 Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, *Climatic Change*, **81**:247-264.
[19] T.J. Santner, B.J. Williams, and W.I. Notz, 2003 *The design and analysis of computer experiments*, Springer.
[20] H. Stommel, 1961, Thermohaline Convection with Two Stable Regimes of Flow, *Tellus*, **13**:224-230.
[21] R. Tokmakian, P. Challenor, and Y. Andianakis, 2012 On the Use of Emulators with Extreme and Highly Nonlinear Geophysical Simulators, *Journal of Atmospheric and Oceanic Technology*, **29**:1704-1715.
[22] I. Vernon, M. Goldstein, and R.G. Bower, 2010 Galaxy Formation: a Bayesian Uncertainty Analysis, *Bayesian Analysis*, **05**(04):619-670.
[23] D. Williamson, M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazak, 2013, History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, *Climate Dynamics*, **41** (7-8), 1703-1729.
[24] K. Zickfield, T. Slawig, and S. Rahmstorf, 2004 A low-order model for the response of the Atlantic thermohaline circulation to climate change, *Ocean Dynamics*, **54**(1):8-26.

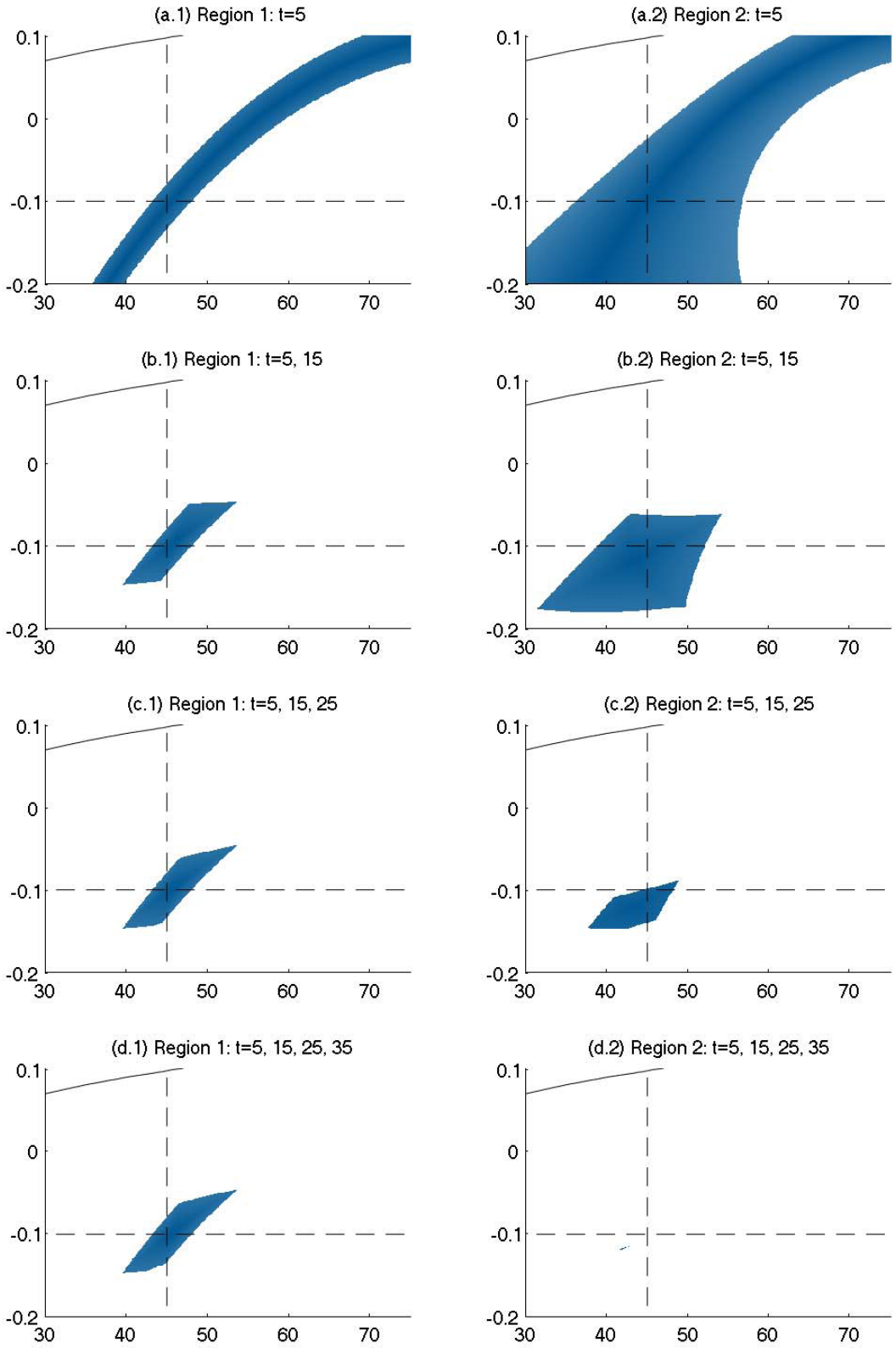

Figure 9. Combined implausibility maps with cut-off at 3 for the emulators in each region at different stages. The black curve represents the boundary between regions 1 and 2 and the dashed lines show the position of the true input value.

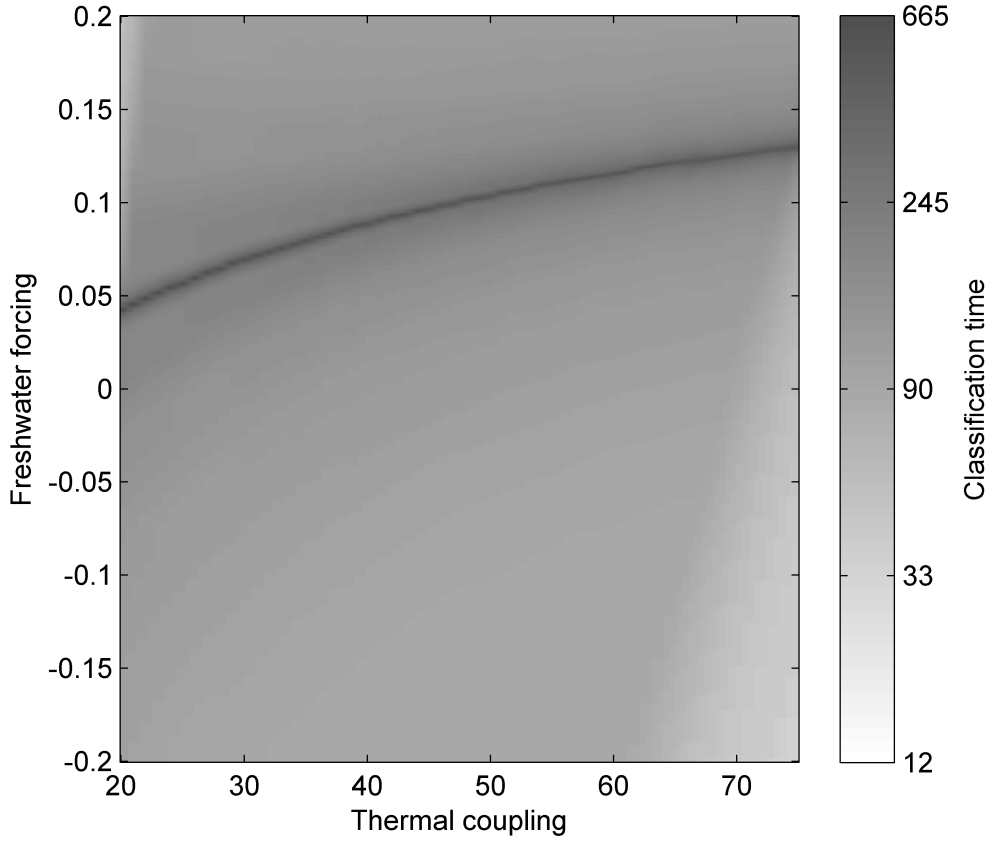

Figure 10. Time t when a region, R_1 or R_2, was ruled out as implausible. A point is in region R_1 if $m_{eq} > 0$ or in R_2 if $m_{eq} \leq 0$.