Precalibrating an intermediate complexity climate model

Journal Item

How to cite:

Edwards, Neil; Cameron, David and Rougier, Jonathan (2011). Precalibrating an intermediate complexity climate model. Climate Dynamics, 37(7-8) pp. 1469–1482.

Climate Dynamics manuscript No.
(will be inserted by the editor)

Precalibrating an intermediate complexity climate model

Neil R Edwards · David Cameron · Jonathan Rougier

Received: date / Accepted: date

Abstract Credible climate predictions require a rational quantification of uncertainty, but full Bayesian calibration requires detailed estimates of prior probability distributions and covariances, which are difficult to obtain in practice. We describe a simplified procedure, termed precalibration, which provides an approximate quantification of uncertainty in climate prediction, and requires only that uncontroversially implausible values of certain inputs and outputs are identified. The method is applied to intermediate-complexity model simulations of the Atlantic meridional overturning circulation (AMOC) and confirms the existence of a cliff-edge catastrophe in freshwater-forcing input space. When uncertainty in 14 further parameters is taken into account, an implausible, AMOC-off, region remains as a robust feature of the model dynamics, but its location is found to depend strongly on values of the other parameters.

Keywords Uncertainty · Probabilistic prediction · Thermohaline circulation · Intermediate complexity climate model

N.R. Edwards
Earth and Environmental Sciences, The Open University, Milton Keynes, MK7 6AA, UK
Tel.: +44 1908 659358, Fax: +44 1908 655151
E-mail: n.r.edwards@open.ac.uk

D. Cameron
Centre for Ecology and Hydrology, Edinburgh, UK

J. Rougier
Department of Mathematics, University of Bristol, BS8 1TW, UK
E-mail: j.c.rougier@bristol.ac.uk

1 Introduction

The credibility of climate predictions rests on the treatment of uncertainty. For a given forcing, uncertainty arises from unknown model error, expressed as the discrepancy between the predicted model state and the actual future climate state. The two most important sources of error in this context are structural error, caused by the imperfect construction of the parameterisations, and parametric error, caused by non-optimal calibration of model parameter values. Errors in initial conditions can be treated as analagous to parametric errors for our purposes.

An archetypal problem is the stability of the Atlantic meridional overturning circulation (AMOC) often equated with the Atlantic thermohaline circulation (THC), although the AMOC forcing is not entirely thermohaline. Changes in the AMOC would have major consequences for European and global climate (Vellinga and Wood, 2002, 2008) but models simulate a wide range of possible future behaviour (Gregory *et al.*, 2005; Stouffer *et al.*, 2006). Most models tend to show a weakening of the present Northern-sinking pattern of AMOC, as measured by the average rate of sinking of water mass in the North Atlantic, in response to anthropogenic carbon emissions. As part of a large-scale comparison of modelling results, Gregory *et al.* (2005) found a 10 to 50% weakening of the AMOC in 140-year simulations with CO_2 increasing to 4 times pre-industrial levels. The AMOC is widely believed to be sensitive to freshwater forcing, either by a stronger hydrological cycle in a warmer climate or by ice-sheet melting, thus much effort has gone into so-called “hosing” experiments in which fresh water is added to the ocean in high northern latitudes. Responses to hosing experiments are also widely spread, with Stouffer *et al.* (2006) finding a reduction between 9 and 62% in response to a not-unreasonable forcing of 0.1 Sv (1 Sv = 10^6 m^3s^{-1}). Synthesising results from 29 simulations performed by 9 separate models for the IPCC’s fourth assessment report (Meehl *et al.*, 2007), weighted by model skill, Schmittner *et al.* (2005) found a weakening of the AMOC by 25 +/- 25% at year 2100.

The prediction of AMOC behaviour thus remains subject to considerable uncertainty, indeed, the thorough elicitation study of Zickfeld *et al.* (2007) revealed that leading experts believe the range of likely behaviour to be considerably wider than that found in models, partly because of known structural deficiencies. The issue of possible overconfidence in such elicitations is covered in the review by Kynn (2008) who argues that any such bias can be expected to be small in well-designed, real-world studies, particularly in predictive situations and where subjects are experienced in making probabilistic judgements.

It is important to realise that model “intercomparisons” do not amount to a quantification of structural model error for three reasons, firstly most studies only consider a set of “best estimate” simulations, thus deliberately avoiding lower probability outcomes and ruling out comprehensive sampling of the distribution. Secondly, the models are usually structurally similar, potentially sharing certain types of error, Thirdly, differences between models will, in practice, be a mixture of structural and parametric components.

A convincing quantification of structural error in AMOC predictions would require quantitative statistical connection between different simulators (Goldstein and Rougier, 2004) and remains some way off. However, parametric error may well be of at least comparable order of magnitude, as evidenced by the wide range of behaviour in single-model ensembles (Edwards and Marsh, 2005; Murphy *et al.*, 2004). Quantification of parametric error requires knowledge of model behaviour throughout a typically high-dimensional parameter space, and thus requires large ensembles of runs. Systematic calibration of models without rigorous quantification of errors can be referred to as tuning. The intermediate complexity C-GOLDSTEIN model (Edwards and Marsh, 2005), part of the GENIE model framework (Lenton *et al.*, 2007), has been used as a test-bed for a range of tuning techniques, firstly by Edwards and Marsh (2005) who used a basic latin hypercube sampling with 1000 simulations, then by Beltran *et al.* (2006) using a cutting-plane optimisation method, Hargreaves *et al.* (2004) with an ensemble Kalman filter, and Price *et al.* (2006) who used a multiobjective genetic algorithm. We will use the same model in this study, but on a different spatial grid (implying previous tuning exercises may not be quantitatively relevant). The process of Bayesian calibration applied to climate models has been described in abstract terms by Rougier (2007), but the practical application would be extremely challenging, even for relatively simple models. The first step in a full calibration is the expert elicitation of prior probability distributions for all important parameters. The expert elicitation of Zickfeld *et al.* (2007) involved full-day interviews with 12 experts, for only a handful of well-studied outputs, but complex models can have hundreds of uncertain inputs. Furthermore, expert elicitation of priors would ideally involve additional quantitative analysis, rather than simple questioning. The second step in Bayesian calibration is a quantification of model behaviour across input space, the final step being the incorporation of constraints from observational data. Using the C-GOLDSTEIN model, Challenor *et al.* (2006) proceeded to the second step in a calibration of AMOC stability and found a surprisingly high probability (around 30 to 40%) of an AMOC collapse by 2100, possibly influenced by the narrow priors, which were largely based on the posterior distributions found in the tuning exercise of Hargreaves *et al.* (2004).

Our objective here is to present an alternative to full calibration that greatly simplifies the procedure, by seeking only to identify simulated outputs which can uncontroversially be classified as unphysical. Our example application, which revisits the issue of AMOC stability in C-GOLDSTEIN, serves to illustrate that even with such weak constraints, statistical modelling of ensembles of simulations can still reveal important features of model behaviour.

2 Precalibration

In this section we start by describing—in general terms—the statistical approach to model calibration, taking into account the imperfection of our model. We contrast this with a ‘lightweight’ alternative that we call ‘pre-calibration’, which makes fewer demands on our judgements. We denote our climate model as $g(\cdot)$. Its inputs $x \in \mathcal{X}$ are those quantities about which we are uncertain: in a climate model these would typically be sub grid-scale parameterisations and flux-corrections. Uncertain initial conditions could be treated similarly in principle, but we do not consider this possibility further here. We refer to $\mathcal{X}$ as the input space, and the set containing $g(x)$ as the output space. The actual value of the climate is denoted y, and the observed climate is denoted z. Here we assume that the selected model outputs correspond to measurable, observable quantities, such that the model error could, in principle, be quantified in terms of the differences $z - y$ and $y - g(x)$.

2.1 Calibration

The inputs to a complex model are often tuned in order to improve the relationship between the model outputs and observations on the underlying system. 'Calibration' is used to describe this process when peformed within a statistical framework; see, e.g., Goldstein and Rougier (2004, 2006), or Rougier (2007) in the context of ensemble-based climate prediction. The standard approach is to assert the existence of some 'best input' x^*, and to quantify the model's structural error in terms of the discrepancy $y - g(x^*)$. The observational errors $z - y$ also need to be quantified, unless they are judged to be dominated by structural error (Rougier, 2007). The probability calculus can then be used, in conjunction with a prior distribution $\Pr(x^*)$, to infer a conditional or posterior distribution $\Pr(x^* \mid z)$: the probability distribution of the best input conditional on the observational data. If a point estimate is needed, e.g. for further evaluations of the model, the value $E(x^* \mid z)$ is a natural candidate. Goldstein and Rougier (2009) discuss the 'best input' approach, and its foundational and practical limitations.

The main challenge with this approach is to quantify the structural error, $y - g(x^*)$. This is an uncertain vector, and, assuming for simplicity that the model is judged to be unbiased and the structural error is chosen to be Gaussian, the quantification of structural error is in terms of a discrepancy variance matrix. This variance matrix is an essential part of the calibration process, and it would be a serious mistake to proceed with the calibration of an imperfect model, such as a climate model, without quantifying it. Ignoring it completely is akin to setting the variance to zero—asserting that the model is perfect except only for uncertainty about the model parameters. This is not acceptable for the current generation of climate models.

Climate scientists have only recently confronted the challenge of specifying the structural error variance (Murphy *et al.*, 2007). Direct attempts are very challenging, thus it is natural to ask whether alternative approaches can be developed which allow for the existence of structural error, and thus do not amount to assuming a model is perfect, but are nevertheless simple enough to be tractable and relatively uncontroversial in their basic assumptions. This is the objective of 'precalibration'. It is less powerful than full calibration, in terms of its ability to provide accurately quantified probabilistic predictions, but it is considerably less demanding and also less subjective. Precalibration does not attempt to quantify structural error as such, but rather to make progress in analysing model behaviour while allowing for the existence of uncertainty and error in general terms.

2.2 Precalibration

The basic idea of precalibration is to rule out some choices of x as candidates for x^*. In order to do this, we begin by identifiying model outcomes that are sufficiently contrary to established system behaviour that they can be relatively uncontroversially classified as 'non-physical'; for example, a pre-industrial Arctic with no sea-ice. If $g(x)$ is judged non-physical, we are prepared to assign a zero or near-zero value to the probability that x is a good candidate for x^*, in other words, we deem x to be an 'implausible' input value. We use the term 'unphysical' to refer to model solutions that disagree strongly with observations rather than to states of the world that could not exist. A collapsed AMOC in a simulation of the modern climate, for instance, will be classed as unphysical, although it could be a physically sensible solution in certain palaeoclimate regimes. Equally, the relevant criteria could, for instance, be biological rather than purely physical.

The attractions of precalibration are: (i) is it based on simple and relatively uncontroversial criteria; (ii) it does not require us to specify a prior distribution for x^*; and (iii) it does not make explicit or detailed use of the actual observations z. Its limitation is that it does not permit us to narrow our set of candidate values for x^* to the extent that a fully-probabilistic calibration using the same evaluations and observations might have done. Nevertheless, in practice there is a balance between the degree to which the ruling-out becomes controversial, and the extent to which the set of candidates for x^* is reduced. Note also that precalibration does not rule out a subsequent calibration using z: there is no double-counting because we do not have to consult z explicitly when classifying certain values for $g(x)$ as non-physical. Ultimately, then, precalibration provides a relatively low-cost opportunity to learn about the model inputs, which does not compromise further analysis.

Ideally, the process of precalibration involves the following two steps.

1. Identify a region in the output space of the model $g(\cdot)$ which is 'non-physical';
2. Map this region back into the input space,

$$\mathcal{N} \triangleq \left\{ x \in \mathcal{X} : g(x) \text{ is non-physical} \right\}. \tag{1}$$

In practice, we cannot compute $g(x)$ for every x. Hence we define the *implausibility* of x, which is the probability that $g(x)$ is non-physical:

$$\mathrm{Imp}(x) \triangleq \Pr(x \in \mathcal{N}) = \Pr(g(x) \text{ is non-physical}). \tag{2}$$

With infinite resouces $\mathrm{Imp}(x)$ would be either 0 or 1, because we would simply evaluate $g(x)$ and see whether or not it is in $\mathcal{N}$. Implausibilities between 0 and 1 arise because in practice we are obliged to predict whether or not $g(x) \in N$, based on an ensemble of model evaluations. Therefore the calculation of the relevant probabilities, and hence of implausibility, is based on an ensemble and on a statistical model. As a result $\mathrm{Imp}(x)$ will not be totally objective, because judgements are involved about where to evaluate the climate model, and how to build the statistical model. With sufficient evaluations the impact of these judgements will be minor, but where resource constraints limit the number of evaluations there will be a trade-off between the transparency of the method, and the additional information supplied through our judgements. In our analysis below we have favoured transparency, but we are fortunate to have a fairly large ensemble (more than a thousand model evaluations). Rougier *et al.* (2009) provides an example of using more detailed judgements about the model.

2.3 Projection

Implausibility scores any point $x \in \mathcal{X}$. However, if $\mathcal{X}$ is not low-dimensional, it is not easy to convey implausibility information. What we would really like to be able to analyse and discuss is the effect of small subsets of the inputs; for example, in our GENIE-I climate model below we would like to be able to identify whether a combination of low values of Atlantic-Pacific moisture flux, `APM`, and high values of atmospheric moisture diffusivity, `AMD`, (Table 1) is likely to be non-physical, and discuss why this might be.

Suppose we are interested in the subset $x_1, ..., x_m$ of the inputs $x_1, ..., x_n$, spanning a subspace $\mathcal{X}_A \subset \mathcal{X}$, where $x_A = (x_1, ..., x_m)$ and $x = (x_1, ..., x_n) = (x_A, x_B)$. We define the projection of implausibility onto the subspace $\mathcal{X}_A$ by asserting that a given point $x_A \in \mathcal{X}_A$ is implausible if for every value of x_B, we expect (x_A, x_B) to be implausible. This notion, first suggested in this context by Craig *et al.* (1997), can be expressed

$$\mathrm{Imp}(x_A) \triangleq \min_{x_B} \mathrm{Imp}(x_A, x_B)\,. \tag{3}$$

If x_A is implausible, i.e., $\mathrm{Imp}(x_A)$ is close to one, then $\mathrm{Imp}(x_A, x_B)$ must be close to one for all x_B, ie all values of x compatible with x_A are likely to be implausible.

To illustrate, imagine that $x = (x_1, x_2)$ and that $\mathrm{Imp}(x)$ is generally low, but has a ridge of high values running along $x_1 = x_2$. In this case, according to (3), both $\mathrm{Imp}(x_1)$ and $\mathrm{Imp}(x_2)$ are low, as we never see the ridge in the one-dimensional projections. But because of the *possibility* of the ridge, it would be wrong to say that all values of x_1 were not-implausible. Therefore an implausible region in a subset of the inputs is strong information, but the absence of such a region does not rule out the possibility of an implausible region in a superset of our subset. In practice, we would hope to find implausible regions in small subsets of the inputs, as these can be visualised graphically.

3 Our climate model

Our climate model, which we denote Genie-I, comprises a reduced physics (frictional geostrophic) 3D ocean model coupled to a 2D energy moisture balance model (EMBM) of the atmosphere and a dynamic-thermodynamic sea-ice model. The ocean model includes realistic bathymetry, an isoneutral and eddy induced mixing scheme and spatially varying drag. The version used in this study is configured on a 64 x 32 grid, with eight logarithmically spaced depth levels in the ocean. In the work here, we use a seasonal version of the model (seasonally varying insolation). See Edwards and Marsh (2005) for a full description of the model. This version of GENIE (also referred to as C-GOLDSTEIN) is orders of magnitude less computationally expensive than most other 3D ocean-climate models, but still retains the nonlinear dynamics of the AMOC, and has thus proven a useful model for demonstration of climate model calibration techniques (Hargreaves *et al.*, 2004; Beltran *et al.*, 2006; Price *et al.*, 2006). However, calibration will depend strongly on the resolution. The previous studies had a lower resolution in longitude, and a constant area for all gridcells, implying different latitudinal distribution of gridpoints with higher equatorial and lower polar resolution. Nevertheless, the choice of input parameter ranges is based on these earlier studies, in particular Edwards and Marsh (2005). In keeping with the philosophy of precalibration, the upper and lower bounds are intended to exclude only uncontroversially extreme values.

The Genie-I inputs are given in Table 1. Many of the inputs are common to other models but some require explanation: the drag parameterisation replaces all nonlinear and diffusive momentum effects with a simple linear friction term for which the inverse coefficient, ODC, has the dimensions of time; this frictional formulation leads to excessive dissipation of momentum, which is countered by a scaling of the windstress by a factor WSF, to give realistic wind-driven flow; the single-layer atmosphere lacks dynamical eddies, thus atmospheric transport is perfomed by diffusion according to fixed latitudinal profile with amplitude AHD and width WAH for heat, and a constant amplitude AMD for moisture. There is also advection by fixed wind fields, scaled by coefficients ZHA and ZMA, but heat is only advected zonally. Modelled amospheric moisture transport from Atlantic to Pacific is relatively weak, but is critical for maintaining the AMOC, so we add a constant Atlantic to Pacific moisture transfer, scaled by the parameter APM. Above a threshold, THP, excess moisture is rained out of the atmosphere instantaneously (in other versions of GENIE, a small timelag is applied). Formally, the time-derivative of velocity is neglected, but at each timestep the calculated velocity is relaxed back to the value at the previous timestep at a rate controlled by parameter LRL.

4 Sequential design

Our intention is to evaluate the parameter-space of Genie-I, in order to identify, if possible, low-dimensional regions that are implausible. These regions will help us to understand Genie-I better, and make our subsequent use of the model more efficient, for example by avoiding model evalutions at implausible input values.

In a pilot study we discovered that the Genie-I solver failed to complete the spin-up at some input values. Such numerical failures could have two possible causes, either the discrete numerical solver has failed to approximate the correct, physically reasonable solution to the continuous model equations, or the solution to the continuous model equations for the given inputs is itself unphysical, featuring extreme values which cause the solver to fail. The distinction between these possibilities may be important for subsequent improvements to the model and solver, but at this stage of the analysis we are concerned with locating implausible input values for a given configuration of the model and solver, thus we treat the failure of the solver to spin-up at x as *prima facie* evidence that $g(x)$ would be non-physical. In other words, $\mathcal{C}^c \subseteq \mathcal{N}$, where $\mathcal{C}$ is that part of the input space where the solver completes, and $\mathcal{C}^c$ is its complement. Further examination (see below) revealed that most of the failures were ultimately physical in origin although in general applications it may not always be practical to determine whether the origin of failure is numerical or physical.

We divided our budget of approximately 2000 evaluations into two parts. In the first part we used a space-filling design over the whole of the input space. We used the result of this ensemble to construct a statistical model for $\Pr(x \in \mathcal{C})$. We find that 341 of the evaluations in this ensemble of 1000 evaluations completed. For the second ensemble we used this statistical model to select evaluations that had a high probability of completion. 799 out of this second ensemble (of 1087) completed. It is important to appreciate that although 2087 evaluations may seem like a lot, they are very sparsely distributed through a 16-dimensional space, which has $2^{16} = 65536$ corners. Despite our ensemble, we remain uncertain about whether $x \in \mathcal{C}$, for an arbitrary $x \in \mathcal{X}$.

We now describe our approach in more detail.

4.1 First design

Design for computer experiments is a well-developed area; see, e.g., the review paper of Koehler and Owen (1996), or the textbook of Santner *et al.* (2003). The standard approach for an initial design is to use a space-filling design such as a maximin Latin

Table 1 Inputs for the Genie-I model. Inputs with a '†' suffix are treated on a logarithmic scale. The standard value of each input is midway between the min and max values (midway between log(min) and log(max) for the logarithmic inputs).

ID	Description	Units	Min	Max
WSF	Windstress scaling factor		1	3
OHD†	Ocean horizontal diffusivity	$10^3\,\mathrm{m}^2\,\mathrm{s}^{-1}$	0.3	3.77
OVD†	Ocean vertical diffusivity	$10^{-6}\,\mathrm{m}^2\,\mathrm{s}^{-1}$	2	200
ODC†	Ocean inverse drag coefficient	days	0.5	5
AHD†	Atmospheric heat diffusivity	$10^6\,\mathrm{m}^2\,\mathrm{s}^{-1}$	1	10
AMD†	Atmospheric moisture diffusivity	$10^6\,\mathrm{m}^2\,\mathrm{s}^{-1}$	0.05	5
WAH	Width of atmospheric heat diffusivity profile	radians	0.5	2
ZHA	Zonal heat advection factor		0	1
ZMA	Zonal and meridional moisture advection factor		0	1
SID†	Sea ice diffusivity	$10^3\,\mathrm{m}^2\,\mathrm{s}^{-1}$	0.3	25
APM	Scaling factor for Atlantic-Pacific moisture flux	Sv	0	0.64
THP	Threshold relative humidity, for precipitation		0.8	0.9
CRF	Climate sensitivity, CO2 radiative forcing	$\mathrm{W\,m}^{-2}$	4.77	8.77
SOC	Solar constant	$10^3\,\mathrm{W\,m}^{-2}$	1.363	1.373
GMR	Greenland melt rate due to global warming	$10^{-3}\,\mathrm{Sv\,degC}^{-1}$	10	30
LRL	Logit of velocity relaxation		3	19

Hypercube. This gives reasonable coverage of the input space, providing good information about the main effect of each input, and some information about the low-order interactions.

A maximin Latin Hypercube treats all of the inputs equally. We make one modification, to prioritise the inputs which we judge to be important, termed the 'active' inputs (Craig *et al.*, 1997, 2001). We identify `OHD`, `AHD`, `AMD`, `WAH`, `ZHA`, and `ZMA` as likely to be active inputs for our evaluations of Genie-I. These were chosen as they control important transports of heat/moisture in the ocean and atmosphere (`ZHA`, `ZMA`: atmospheric advection; `AHD`, `AMD`: atmospheric heat and moisture diffusion; `OHD`: ocean heat diffusion). We would like our design to be sensitive to interactions among these inputs in particular. Therefore, having generated a 1000×16 maximin Latin Hypercube, we examine all $\binom{16}{6} = 8008$ sets of six columns, to find the set with the best properties for identifying interactions. We quantify this using the determinant of the 6×6 correlation matrix. We assign our six active inputs to the six columns with the largest determinant; crudely, if there was a linear combination among the columns the determinant would be zero, and this is the kind of design we would like to avoid. This type of assignment of inputs to columns is a simple way to prioritise some of the inputs on the basis of weak judgements about which inputs will be active. Bayesian experimental design (see, e.g., Chaloner and Verdinelli, 1995) allows for more detailed judgements, where they exist. Note that while the choice of active inputs may be more controversial than other choices in the precalibration process, it can be verified *a postiori*, see Section 6.1, and is designed purely to aid the statistical modelling process. The conclusions should not be significantly affected.

4.2 Modelling the probability of completion

We evaluate Genie-I at the 1000 values for x, of which 341 complete. We would like to map the relationship between x and completion, in order to avoid performing evaluations with a high chance of failing to complete in the second part of our experiment. For simplicity and transparency, we use standard statistical tools for this task, namely logistic regression with stepwise variable selection, implemented in the Statistical Computing Environment `R` (R Development Core Team, 2004), using the `stepAIC` function (in the `MASS` library, see Venables and Ripley, 2002). There are some technical concerns about applying logistic regression to the output of a deterministic model such as Genie-I (discussed in Rougier *et al.*, 2009), but we do not consider these to be critical for what is effectively an exploratory analysis.

First, we transform the inputs `OHD`, `OVD`, `ODC`, `AHD`, `AMD`, and `SID`, by taking logarithms. Then we map all inputs onto the range $[-1, 1]$ using the minimum and maximum values in Table 1. This range makes odd and even functions orthogonal with respect to a uniform weighting function, improving the selection of terms in the stepwise selection. We initialise our statistical model with a constant and linear terms only. Then we grow the statistical model using stepwise selection on all quadratics, cubics, and two- and three-way interactions (see, e.g., Draper and Smith, 1998, ch. 15). Our chosen statistical model maximises the Akaike Information Criterion (AIC).

Fifty-five terms are added using this approach, and the linear term in `LRL` is deleted (indicating that `LRL` has little influence on completion), so that there are no terms in `LRL` in the resulting statistical model; the `SOC` input is also marginal. The first interactions selected (i.e. most important) are `WAH:AHD`, `ZMA:OHD`, `ODC:OHD`, `AMD:AHD`, `AMD:OHD`,

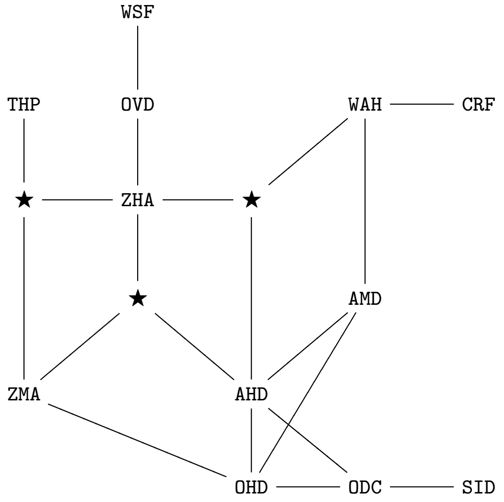

Fig. 1 Graph of the main relationships between the inputs for determining the probability of completion. An edge between two inputs indicates a two-way interaction. Three edges to a star indicate a three-way interaction and all three two-way interactions.

ZHA:AHD, ZMA:AHD, and ZHA:WAH. In Figure 1 we present a simple visual summary of the way in which the inputs interact with each other. We construct an undirected graph where the vertices are the inputs, and edges indicate interactions. We do not show all the interactions, since that would be hard to read, instead we show the top interactions according to the order in which they are selected. In the absence of a thorough sensitivity analysis of the form of the graph to the details of the statistical model fitting process, the graph must be interpreted with great caution. Nevertheless, where parameters are multiply connected, this suggests that they are relatively important in the determination of completion, and where parameters are linked, there may be nonlinear interactions which are also important. Conversely, parameters which are isolated or do not appear at all may have relatively little influence.

The completion graph can be interpreted in terms of the analysis of failure modes. In an analysis of 100 randomly selected failed simulations, 98 failed apparently as a result of extremely low temperatures, below -150°C. Of these, 12 had high values of AHD and WAH, apparently leading to numerical failure via diffusional instability in the atmosphere. All but 18 of the remaining failures appeared to result from insufficient atmospheric heat transport to the poles, with low values of some or all of the parameters WAH, AHD, AMD and ZHA. In the graph, a high-diffusion failure mode involving WAH, AHD, is visible around the upper-right star, this region of the graph also contains a low-diffusion failure mode involving WAH, AHD, AMD (which implies latent heat transport through moisture transport) and ZHA, the latter two having no direct connection, perhaps because zonal heat advection can only act on poleward heat advection indirectly, via zonal redistribution of heat, eg between land and ocean regions. Such nonlinear effects connecting atmosphere and ocean (via ZMA - OHD) and involving heat and moisture fields, appear in the lower left of the graph. Apart from this link, ocean parameters are surprisingly isolated at the top and bottom of the graph, suggestive of a relatively weak influence on completion. This may be related to a better initial constraint on ocean parameters, or the better conservation of properties in the ocean part of the coupled system (where heat is conserved in the interior), or a less heavily parameterised model than the simple EMBM atmosphere, or simply a better solver, and hence a lesser role in failures. There is no obvious evidence for high fluid-velocity Courant-Friedrichs-Lewy (CFL) failures (eg near-limiting velocities prior to numerical failure), and this failure mode was not identified as important, again probably reflecting conservative input parameter ranges.

As a form of statistical model criticism, we can use the resulting statistical model to compute a point prediction for $\Pr(x \in \mathcal{C})$ at any $x \in \mathcal{X}$. As a simple guide to the quality of our statistical model, the following table shows the predicted and actual outcomes for the ensemble, based on our statistical model and a threshold of 50%:

	$x \in \mathcal{C}^c$	$x \in \mathcal{C}$	Sum
$\Pr(x \in \mathcal{C}) < 0.5$	617	40	657
$\Pr(x \in \mathcal{C}) \geq 0.5$	42	301	343
Sum	659	341	1000

(4)

This shows a misclassification error for acceptance, defined as the probability that a point above our threshold fails to complete, of $42/343 \approx 12\%$, and a misclassification error for rejection, defined as the probability that a below-threshold point would have completed, of $40/657 \approx 6\%$. These are much better than could be achieved from a more limited knowledge of the ensemble. The case of no predictive knowledge other than the mean, for example, analagous to tossing a biased coin with probabilities 341/1000 and 659/1000, would give misclassification errors of 66% and 34% for acceptance and rejection respectively.

4.3 Second design

We use our statistical model for $\Pr(x \in \mathcal{C})$ to assess each candidate for our second design. We set a threshold ν and keep the candidate x if $\Pr(x \in \mathcal{C}) \geq \nu$. There are two errors we can make with this approach. First, we can screen out an x which would have completed. Second, we can fail to screen out an x which does not complete. As ν decreases from one to zero to one we trade the probability of the first error (which is one when $\nu = 1$) against the probability of the second (which is one when $\nu = 0$). Where we set ν will depend on the cost of the two types of error. We regard the the first error as the more critical, and we aim to choose a value for ν that makes the first error roughly half as probable as the second. As shown in the table in (4) the choice of $\nu = 0.5$ satisfies this criterion, based on the results of the first ensemble. About 34% of the evaluations get past the threshold, so if we generate an initial design of $1000/0.343 \approx 2915$ over the whole of $\mathcal{X}$ then after screening we should end up with about 1000 evaluations in our second design, favouring $\mathcal{C}$.

We follow the same steps as before, generating a 2915×16 maximin Latin Hypercube, and assigning the active inputs to the best subset of six columns. Then we predict $\Pr(x \in \mathcal{C})$ for each candidate value for x in turn, and keep those for which this is no less than 0.5. The result is 1087 evaluations in the second ensemble. After evaluating them, we find that 799 complete, or 74%.

4.4 Transient runs

At this stage of the experiment, we have 2087 evaluations, of which 1140 complete their spin-up. We now run each spun-up evaluation forward using one percent per annum compound increase in CO_2 from 1850 to 2100: in the case of Genie-I this is represented as a direct increase in radiative forcing. At this stage we lose another 94 evaluations (39 from the first design and 55 from the second), for which the solver failed to handle the transient behaviour; again, we classify these as non-completers. This leaves us with 1046 completed evaluations after both the spin-up and transient phases.

5 Implausibility analysis

5.1 Non-physical ranges

The precalibration outputs and ranges for the Genie-I model are given in Table 2. Note the deliberately wide 'physical' ranges. We determined these limits by considering what we would class as non-physical for our Genie-I model. Although we treated the five target outputs individually, it turns out that there is a dominant non-physical mode, which is the absence of positive AMOC cell. In this case the maximum Atlantic streamfunction will be too low; the temperature in the upper Atlantic will be too low; and the Atlantic will be too fresh relative to the Pacific, as the interbasin salinity contrast is known to be closely associated with the northern-sinking positive AMOC state, presumably because denser, high-salinity water is prone to sink in the North Atlantic. Table 2 also shows the percentage of evaluations in our ensemble that are too low or too high in at least one of the years 1850, 1900, 1950, 2000. In total, 23% of our 2087 evaluations satisfy all five ranges, which is to say that 77% are classified as non-physical.

5.2 Statistical modelling

We now focus on a second set of probabilities, namely $\operatorname{Imp}(x)$, as defined in section 2.2. Rather than construct a single statistical model, we choose to construct two, and combine them using the rules of probability:

$$\begin{aligned}\operatorname{Imp}(x) &= \Pr(x \in \mathcal{N}) \\ &= 1 - \Pr\bigl(x \in N^c\bigr) \\ &= 1 - \Pr\bigl(x \in \mathcal{N}^c,\, x \in \mathcal{C}\bigr) \\ &= 1 - \Pr\bigl(x \in \mathcal{N}^c \mid x \in \mathcal{C}\bigr)\ \Pr(x \in \mathcal{C}) \end{aligned} \tag{5}$$

where the introduction of $x \in \mathcal{C}$ in the third line follows from $x \in \mathcal{N}^c \implies x \in \mathcal{C}$. and '|' denotes 'conditional upon'. The last line follows from the definition of conditional probability. This decomposition allows us to construct the full implausibility from our model of completion and from the ensemble of completed runs.

The statistical model for $\Pr(x \in \mathcal{C})$ is similar to the statistical model we have already constructed from the first part of our design (see section 4.2). We refit the statistical model, with the same choice of regressors, but now using the full ensemble of 2087 evaluations. The incomplete evaluations in the spin-up of the second ensemble

Table 2 Precalibration ranges for the climate values corresponding to the Genie-I outputs. The final two columns show the percentage of completed runs that lie outside the range. The separation between upper and deep water is at 1158 m depth, the Southern Ocean includes all point south of the tip of South America, Atl. and Pac. sectors include all points in these sectors north of the Southern Ocean.

Climate quantity	Units	'Physical'		% too	
		min –	max	low	high
Max. Atl. streamfunction	Sv	10 –	35	24	31
Mean temp. in the upper Atl.	°C	6 –	12	6	7
Mean temp. in the deep Atl.	°C	2 –	8	22	9
Diff. in mean salinity between the upper Atl. and the upper Pac.	PSU	0 –	1.5	22	2
Diff. in mean temp. between the upper Atl. and the upper Southern Ocean	°C	−1 –	9	17	0

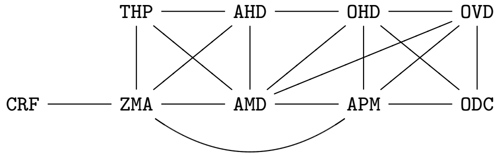

Fig. 2 Graph of the main interactions between the inputs for determining the probability of a not-unphysical output, among evaluations that complete. See the caption to Figure 1 for details.

are likely to be particularly informative, because they contradict the prediction of the model fitted on the first ensemble alone. The misclassification rate rises to 15%, but a rise is to be expected because we do not re-select the regressors in the model, as a precaution against over-fitting. If the mechanism that triggers a solver failure in the transient phase were different from that in the spin-up, it would tend to cause a rise in the misclassification rate, but we have no evidence that this has occurred in our case.

The statistical model for $\Pr(x \in \mathcal{N}^c \mid x \in \mathcal{C})$ is fitted only on the 1046 evaluations which complete, in the same way as described in Section 4.2. The misclassification rate of the statistical model is 0.5%. Figure 2 shows the graph of the main relationships between the inputs, after building our statistical model. Perhaps not surprisingly, this graph is easier to interpret than the graph for simulation failures. There is a broad separation between ocean parameters on the right and atmosphere parameters on the left, with parameters in the centre of the graph being of the greatest significance for ocean-atmosphere interaction and exhibiting the largest number of interactions in the graph, six for AMD, five for APM and OHD. Ignoring CRF the lower left region comprising THP, ZMA, AMD and APM all control moisture flux, whereas the upper right four parameters control heat flux. The graph reveals which parameters are most important in ocean-atmosphere interactions controlling the AMOC (the principal unphysical mode) and confirms the importance, but relative isolation, of ocean drag (ODC), and of the precipitation threshold (THP) and moisture advection (ZMA) in the atmosphere, parameters which it can be tempting to ignore in trying to understand the model.

We compute the implausibility using two statistical models combined, rather than just one (which we could have constructed using the 481 not non-physical outcomes from 2087 evaluations), because this allows us to attribute high implausibility consistently between the two possible causes: a failure to complete at x or, if complete, a non-physical outcome. An additional advantage is that the statistical model for $\Pr(x \in \mathcal{N}^c \mid x \in \mathcal{C})$ is more focused than the model for $\Pr(x \in \mathcal{N}^c)$, being constrained to a smaller region of the input space, and being descriptive of a simpler outcome. This makes it easier to fit the statistical model (*cf* the low misclassification rate), and—we hope—easier to interpret the result.

6 Further analysis using implausibility

At this stage we have derived a statistical model for $\mathrm{Imp}(x)$ for all values of the input vector x in our input space. This function is many orders of magnitude cheaper to evaluate than the original numerical model, but its form, as a multidimensional function of its inputs, potentially contains valuable information about the behaviour of the underlying model. To illustrate how the implausibility function may be interrogated to obtain such information, we now consider three linked examples. First we order the parameters by importance, then we project the implausibility onto the four most important parameters, then we turn to the existence of the cliff-edge AMOC catastrophe.

6.1 What are the important inputs?

A simple scalar measure can be used to summarise the importance of each input in determining implausibility. Here, an input is deemed important if it can cause a large change in implausibility. Note that this differs from the more usual interpretation, in which an 'important' input is one which can cause a large change in $g(x)$, as identified in a sensitivity analysis. Therefore for each input in turn we take a sequence of values from small to large, and for each value we compute the implausibility over a space-filling design in the other inputs. We then take the mean absolute value for the changes in these implausibilities as the value increases, and summarise these in a single mean value for each input.

The result is shown in Figure 3. The two inputs `AHD` and `AMD` are the most important, followed by `WAH` and `OHD`. The first five inputs were among the six specified as 'active' inputs in our experimental design, providing an *a postiori* verification of their importance. Indeed the ordering suggests that the quantitative importance of inputs in controlling implausibility is primarily determined by their effect on meridional heat and moisture transport. Note that the effect of each input is measured relative to its assumed input range, in other words to our uncertainty about its best input value. In the heavily parameterised, largely diffusive EMBM atmosphere of Genie-I, the weakly bounded diffusivity amplitudes, `AHD` and `AMD` which strongly control heat and moisture transport, thus appear as the dominant parameters. The next six inputs also play significant roles in global heat or moisture transport, the ocean drag coefficient `ODC` by exerting a frictional drag on the large-scale ocean transport, and the ocean vertical diffusivity `OVD` via its effect on the AMOC. The remaining eight parameters generally have only indirect effects on global-scale transports, with the exception of `ZHA` which was amongst our six 'active' inputs but, unlike `ZMA`, does not affect meridional transport, and furthermore is constrained to a small maximum value relative to the diffusive transports, possibly explaining its relatively minor role in implausibility.

6.2 Implausibility of the four most important inputs

We now project implausibility onto the four most important inputs identified in Section 6.1. Figure 4 shows a four-way layout. The lightest areas have implausibility of less than 5%, while the darkest areas have implausibility of greater than 95%. The difference between the left- and right-hand panels shows that low values of `WAH` are more implausible than high values, and the lack of difference between the top and bottom panels shows that changing `OHD` does not alter this. Within the right-hand panels, very large values of `AHD` are implausible for all values of `AMD`. As `AHD` and `WAH` both affect the form of the atmospheric thermal diffusivity as a function of latitude, implausibility at high `AHD` and `WAH` could be related to high-diffusivity numerical breakdown. Nevertheless, even though `AHD` and `WAH` are very closely related, their effects on implausibility are not trivially related. The lowest values of all four transports, in the bottom left of

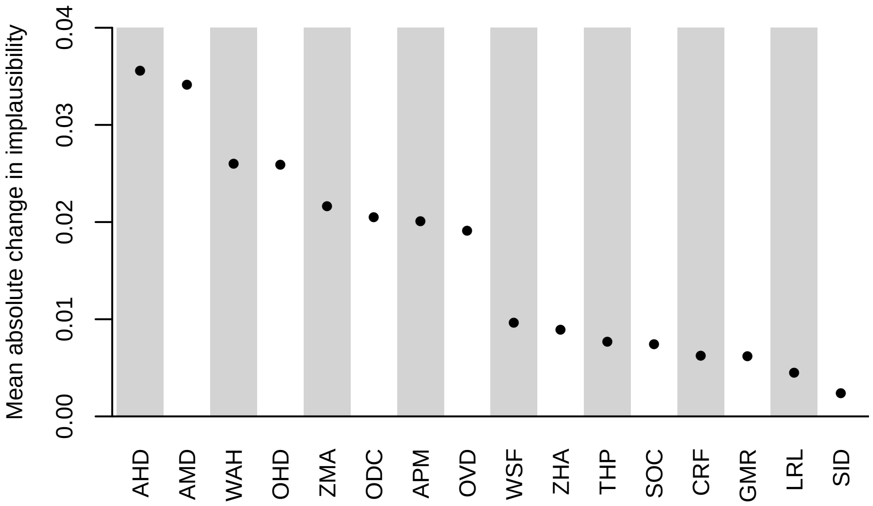

Fig. 3 Scalar summary of the importance of each input in determining implausibility, ordered from most to least important (see text for details). The value indicates the degree to which a change in the value of the input changes implausibility over the input space as a whole.

the upper left plot, also show high implausibility, possibly related to the unphysical polar conditions identified previously for low diffusion. The interaction between heat and moisture diffusivities `AHD` and `AMD` is not simple: starting from the saddle point in the lower left plot, a reduction in `AMD` increases implausibility but can be offset by either an increase or a decrease in `AHD` or, to a lesser extent, a decrease in ocean heat diffusivity `OHD`. It could be relevant that increased moisture transport implies increased latent heat transport but, on the other hand, meridional moisture and heat transport have opposing direct density effects on driving the thermohaline circulation. Alternatively, the nonlinear features of the plot may be related to competition between the five different physicality targets. We do not attempt to rationalise the form of the implausibility surface in any more detail, since our objective was simply to illustrate the potential for mapping out its behaviour in multiple dimensions. In general, the surface will have some complicated dependence on all 16 inputs. In the next section, we focus on a more tractable projection onto only two dimensions.

6.3 The AMOC 'cliff-edge' catastrophe

We now consider the question of the existence of a 'cliff-edge' AMOC catasotrophe in freshwater forcing input space, as identified by Marsh *et al.* (2004), by considering projections of implausibility onto relevant subspaces of the inputs. Figure 5 compares a cross-section through $\mathrm{Imp}(x)$ with the relevant projection $\mathrm{Imp}(x_A)$ from (3). In the left-hand panel of Figure 5 `APM` and `AMD` have been varied in a grid, with the other 14 model inputs held fixed at their standard values. This picture tells us about GENIE-I's response on one 2-dimensional plane through the 16-dimensional model input space. The 'cliff-edge' indicates that on this plane there is a sharp division between settings

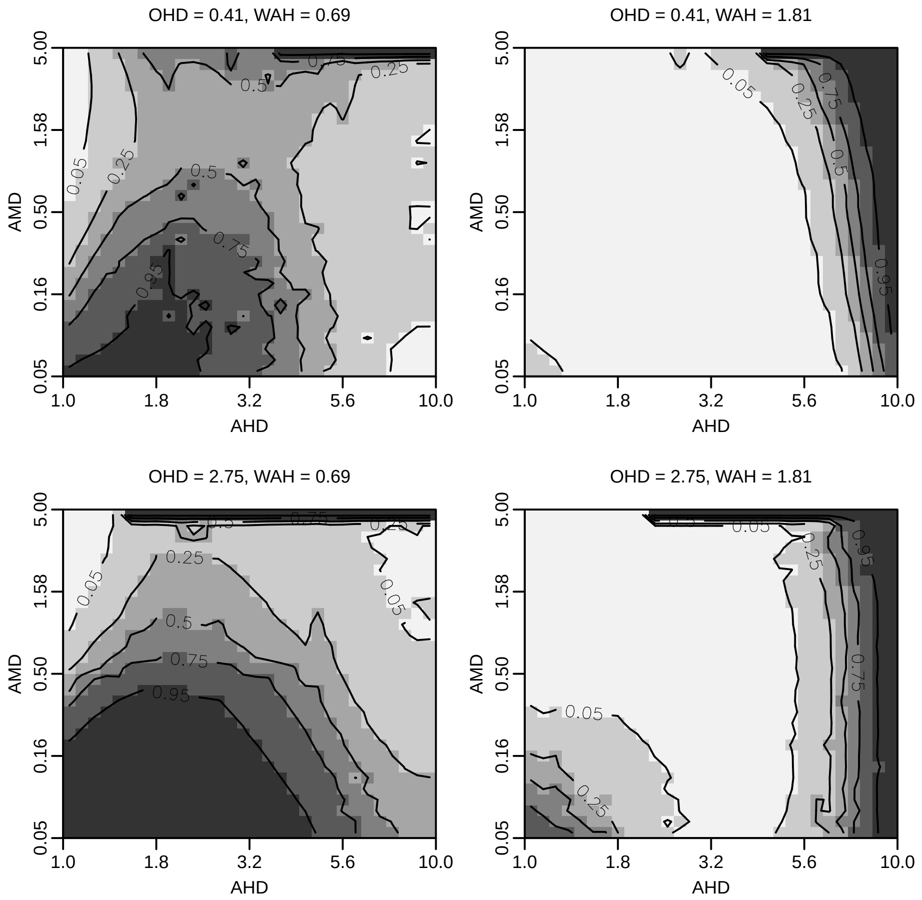

Fig. 4 Implausibility projected onto the four most important inputs, as judged from Figure 3; the darker areas are more implausible, and the contour lines are at 5%, 25%, 50%, 75%, and 95%. Each panel shows `AHD` and `AMD`, while the four panels comprise a two-way layout of `OHD` (top low, bottom high) and `WAH` (left low, right high). Note that both `AHD` and `AMD` are modelled on a logarithmic scale. Units are given in Table 1.

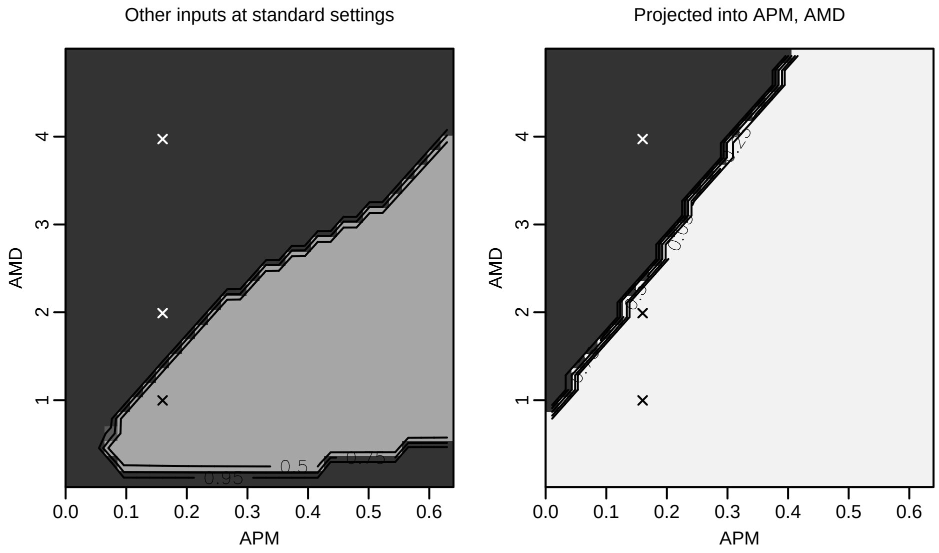

Fig. 5 The probability that the model output is non-physical (Imp(x)) shown for combinations of the Atlantic-Pacific moisture flux, `APM` (Sv), and the atmospheric moisture diffusivity, `AMD` ($\times 10^6 \mathrm{m}^2\mathrm{s}^{-1}$). (a) All other model inputs set to their standard values, see Table 1. (b) Implausibility, projected through the other model inputs, using eq. (3). Darker shading indicates a larger probability. Crosses indicate the points plotted in Figure 6.

while the different resolution and lack of seasonality could also have a bearing on the failure modes.

The lefthand panel of Figure 5 tells us nothing about the model input space as a whole. The righthand panel, on the other hand, does exactly this, as it shows the *projected implausibility*, for `APM` and `AMD`, which involves projecting through the other 14 model inputs. Minimising over the other 14 model inputs cannot result in an implausibility that is larger than that when the other 14 are at the standard values, hence no point in the righthand panel can be darker than in the lefthand panel. The result of projection is that much of the implausible region disappears: for low moisture diffusivity `AMD` and high Atlantic-Pacific moisture flux `APM`, compensating adjustments in other parameters can give rise to physical model output. On the other hand, the low `APM`, high `AMD` region, corresponding to the AMOC cliff-edge, shifts towards more extreme values, but otherwise remains intact. In this region, even wide-ranging adjustments in 14 other parameters apparently cannot produce physical output.

To illustrate the connection between the cliff-edge and the AMOC, Figure 6 shows the AMOC in three simulations corresponding to the crosses marked on Figure 5 along a transect across the cliff-edge. At each point, the values of the remaining 14 inputs are chosen to minimise the implausibility $\text{Imp}(x)$. As expected, in the uppermost panel, corresponding to the implausible region in the projected `APM-AMD` space, the AMOC is in a fully collapsed state. The middle panel represents an intermediate point on the cliff edge itself, at which the least implausible state, as shown, has a visible, but very weak positive AMOC cell in the deep Atlantic. The lower panel shows a location which is plausible even at standard values of the remaining parameters, where the least implausible inputs give a strong positive AMOC.

Note that the Figures 5b and 4 involve non-trivial computation, as the calculation of $\text{Imp}(x)$, from (3), requires a numerical minimisation of the statistical model for $\text{Imp}(x)$ over all the input dimensions not shown in the figures. The projection code divides the inputs into three types: the ones we are projecting onto, other active inputs, and remaining inputs. The 'other active' inputs are explicitly minimised over, while the remaining inputs are spanned with a space-filling design, (the Sobol sequence, implemented in Würtz, 2007). The minimum over the points in the space-filling design is taken to be the minimum over the whole input space. Therefore our implausibility values are upper bounds, but sensitvity tests suggest that our results are relatively accurate.

7 Summary and discussion

Perturbed physics experiments (PPEs) allow us to account for our uncertainty about the values of the inputs to a complex model, such as an EMIC. Typically we express our uncertainty marginally, input-by-input, for example in terms of ranges and simple transformations, as we have done in our example (Table 1). A problem can arise in this type of experiment: the model's solver might break down at some combinations of input values. Typically the solver will be tuned to perform well in the region centred on the model's standard input values. It may also be robust against one-input perturbations; e.g. axial designs in which each input in turn is taken to its minimum and maximum values, with all other inputs at their standard values (see, e.g., Murphy *et al.*, 2004). but it may break down when several inputs are varied simultaneously.

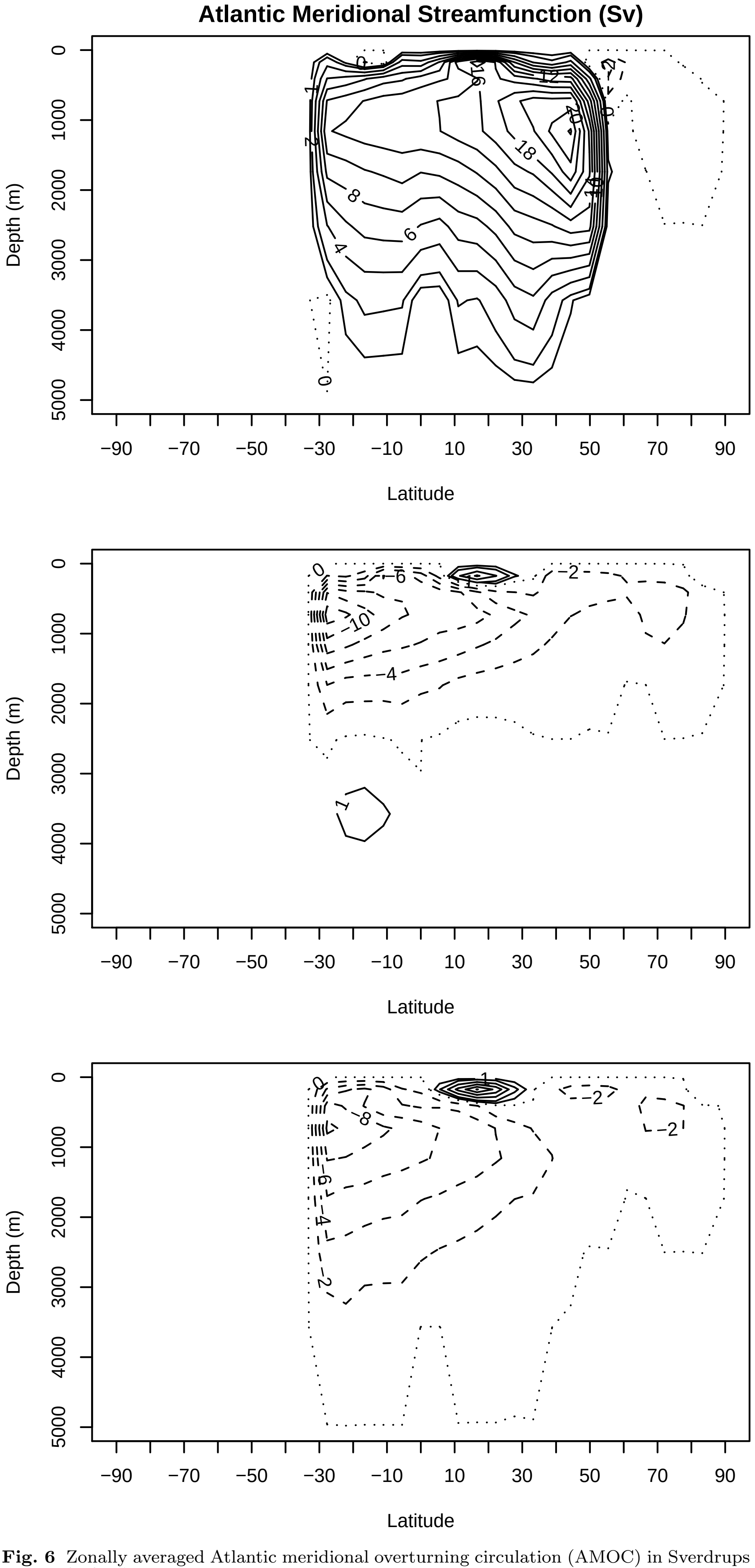

Fig. 6 Zonally averaged Atlantic meridional overturning circulation (AMOC) in Sverdrups (1 Sv = $10^6 \mathrm{m}^3\mathrm{s}^{-1}$) for three simulations corresponding (in vertical order) to the least implausible inputs at the three points indicated as crosses on Figure 5. Dashed lines correspond to negative values, latitude is in degrees.

This is exactly the problem we faced with our Genie-I EMIC, where combinations of extreme (and even not-so-extreme) input values caused the model to fail to complete its spin-up. In this situation we can write a more robust solver (e.g. take smaller time-steps or make more fundamental changes to the model), or we can treat the solver failure as informative for the model. After investigation, we adopted the latter course, and classified those input values for which the solver failed as *prima facie* implausible (for this particular model setup). This was a particularly convenient choice in our analysis, but it is also a natural generalisation of the current practice of only running complex models at their standard input values, which amounts to treating all non-standard choices of the input values as implausible (i.e. not worth evaluating). Our approach is a generalisation because we treat the standard value as only one point within a set of not-implausible input values. Our approach is best understood from the standpoint of calibration, which attempts to find the 'best input' value x^*. Precalibration is concerned with reducing the set of possible candidates for x^*. In either case, we must begin by fixing a definition of our model and its solver, and deciding which parameters are available as inputs. These choices could be revisited if solver failure turns out to be a major issue, as indeed could the form of the parameterisations themselves. Indeed, learning about model parametric error would ideally constitute part of an iterative process to modify both solver design and model parameterisations. The treatment of non-completions is liable to be even more important in more expensive models and alternative approaches could be envisaged, such as including timestep length as a variable parameter. In any case, it will be desirable to avoid excessive non-completions, which are largely wasted simulations.

One of the difficulties of PPEs is that it can be hard to specify our prior uncertainty about the best value of the model inputs. This is often because of difficulties with the operational definition of the model inputs, a problem that becomes more acute in lower-resolution models. Ideally, we would have sufficient observations that, in a statistical calibration, our quantification of prior uncertainty would be relatively unimportant; we could then use wide intervals and simple distributional shapes (e.g., triangular, Beta, Gamma). Unfortunately, this is seldom the case with climate models, where the observations, though abundant, are strongly correlated, so that the likelihood function, ie the region of "good" inputs to the model, tends not to concentrate, but to have long ridges (Rougier, 2007). Another problem is that this calibration requires us to quantify a measure of our model's structural error: this is very challenging.

In this paper we have proposed a simpler version of calibration, which we term *precalibration*, based on the notion of *implausibility* (Craig *et al.*, 1997). Precalibration is a low-cost way of ruling out input values that give rise to non-physical outcomes, and requires us only to specify what outcomes we deem to be non-physical. We use our ensemble to construct a statistical model that allows us to compute the probability that any particular input value will give rise to a non-physical outcome. It is important in this case that our ensemble explores the model's input space in an efficient way, so that we get as much information as possible from our finite set of evaluations. In this paper we have used space-filling designs from the statistical field of Computer Experiments, and we have used sequential design to avoid evaluations likely to be non-physical.

The extent to which the physicality criteria are uncontroversial will, in practice, be a compromise against the extent to which the candidate region for x^* is reduced. Tighter bounds of physicality imposed on the model output would, in general, reduce the size of the region of not-implausible inputs, but make the ruling out process correspondingly more controversial. Similarly, although prior distributions for inputs are not required, narrower input ranges may lead to better resolution of the output space, but would imply more controversial *a priori* decisions. In principle, however, the objective is only to remove regions with zero probability (which implies that precalibration should not distort any subsequent calibration). This may be highly pertinent in probabilistic risk assessments which are driven by the tails, such that 'almost implausible' inputs are associated with high costs. Multiple iterations of precalibration (which may either increase or decrease the implausible region) could be highly valuable in such caseses since the exercise focuses implicitly on defining the edges of acceptable space. To proceed to a probabilistic risk analysis, however, requires explicit weighting of outcomes.

In our illustration with the Genie-I EMIC we have used implausibility to identify implausible choices for various selected inputs. In so-doing we have generalised the analysis of Marsh *et al.* (2004), which considered `APM` and `AMD` only, and we have shown that, in our model, the existence of a cliff-edge catastrophe is robust to the inclusion of uncertainty about more model inputs, but that the location of the cliff-edge depends strongly on other parameters. It is worth stressing that our analysis uses an ensemble of model evaluations which is completely general; which is to say that many other questions can also be addressed using the same ensemble. Given that ensembles are expensive and time-consuming to generate, we would strongly recommend the use of statistical experimental design techniques to construct general purpose ensembles. These can then be used to address specific questions using the techniques we have outlined here. As an example, Holden *et al.* (2009) apply precalibration to the estimation of glacial and future climate sensitivity and changes in terrestrial carbon storage. Their analysis demonstrates that the application of weak constraints on model inputs and outputs, even in two contrasting climate states, still allows for a wide range of predicted behaviour. For a detailed analysis, more statistically-intensive approaches are also possible (see, e.g. O'Hagan, 2006; Rougier and Sexton, 2007; Rougier, 2008). However, these require more specialised statistical input and more computing resources. But an initial exploratory analysis using implausibility is inexpensive and may often prove fruitful.

Acknowledgements The authors gratefully acknowledge the help of Andrew Price in performing large ensemble simulations, as well as support from the UK Natural Environment Research Council (NERC) RAPID climate change programme (via the THCMIP project), the NERC QUEST programme and the UK Department for Environment, Food and Rural Affairs (via the UK Met Office). Part of this work was completed while Rougier was a Duke University Fellow at the Statistical and Applied Mathematical Sciences Institute (SAMSI), Durham NC.

References

A. Agresti, 2002. *Categorical Data*. New York: John Wiley & Sons, second edition.

C. Beltran, N.R. Edwards, A. Haurie, J.-P. Vial and D.S. Zachary, 2006. Oracle-Based Optimization Applied to Climate Model Calibration. *Environmental Modelling and Assessment*, **11**, 31–43.

P.G. Challenor, R.K.S. Hankin and R. Marsh (2006) Towards the Probability of rapid Climate Change. in *Avoiding Dangerous Climate Change* Eds. Schellnhuber, H.J., W. Cramer, N. Nakicenovic, T. Wigley amd G Yohe. Cambridge University Press 53–63.

K. Chaloner and I. Verdinelli, 1995. Bayesian experimental design: A review. *Statistical Science*, **10**(3), 273–304.

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

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments. In C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi, and N.D. Singpurwalla, editors, *Case Studies in Bayesian Statistics III*, pages 37–87. New York: Springer-Verlag. With discussion.

N.R. Draper and H. Smith, 1998. *Applied Regression Analysis*. New York: John Wiley & Sons, 3rd edition.

N.R. Edwards and R. Marsh, 2005. Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model. *Climate Dynamics*, **24**, 415–433.

M. Goldstein and J.C. Rougier, 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, **26**(2), 467–487.

M. Goldstein and J.C. Rougier, 2006. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, **101**, 1132–1143.

M. Goldstein and J.C. Rougier, 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, **139**, 1221–1239. With discussion.

J.M. Gregory, K.W. Dixon, R.J. Stouffer, A.J. Weaver, E. Driesschaert, M. Eby, T. Fichefet, H. Hasumi, A. Hu, J.H. Jungclaus, I.V. Kamenkovich, A. Levermann, M. Montoya, S. Murakami, S. Nawrath, A. Oka, A.P. Sokolov and R.B. Thorpe, 2005. A model intercomparison of changes in the Atlantic thermohaline circulation in response to increasing atmospheric CO2 concentration. *Geophysical Research Letters*, **32**, Art. No. L12703.

J.C. Hargreaves, J.D. Annan, N.R. Edwards and R. Marsh, 2004. Climate forecasting using an intermediate complexity Earth System Model and the Ensemble Kalman Filter. *Climate Dynamics*, **23**, 745–760.

P.B. Holden, N.R. Edwards, K.I.C. Oliver, T.M. Lenton and R.D. Wilkinson 2009. A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. *Climate Dynamics*, DOI 10.1007/s00382-009-0630-8.

J.R. Koehler and A.B. Owen, 1996. Computer experiments. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics, 13: Design and Analysis of Experiments*, pages 261–308. North-Holland: Amsterdam.

M. Kynn, 2008. The 'heuristics and biases' bias in expert elicitation. *J. R. Statistical Soc. A*, **171**, 239–264.

T. M. Lenton, R. Marsh, A. R. Price, D. J. Lunt, Y. Aksenov, J. D. Annan, T. Cooper-Chadwick, S. J. Cox, N. R. Edwards, S. Goswami, J. C. Hargreaves, P. P. Harris, Z. Jiao, V. N. Livina, A. J. Payne, I. C. Rutt, J. G. Shepherd, P. J. Valdes, G. Williams, M. S. Williamson and A. Yool, 2007. Effects of atmospheric dynamics and ocean resolution on bi-stability of the thermohaline circulation examined using the Grid ENabled Integrated Earth system modelling (GENIE) framework. *Climate Dynamics*, **29**, 591–613.

R. Marsh, A. Yool, T.M. Lenton, M.Y. Gulamali, N.R. Edwards, J.G. Shepherd, M. Krznaric, S. Newhouse, and S.J. Cox, 2004. Bistability of the thermohaline circulation identified through comprehensive 2-parameter sweeps of an efficient climate model. *Climate Dynamics*, **23**, 761–777.

G.A. Meehl *et al.*, 2007. in Climate Change 2007: The Physical Science Basis (eds Solomon, S. et al.) Ch. 10 Cambridge.

J.M. Murphy, B.B.B. Booth, M. Collins, G.R. Harris, D.M.H. Sexton, and M.J. Webb, 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1993–2028.

J.M. Murphy, D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins, and D.A. Stainforth, 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772.

A. O'Hagan, 2006. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, **91**, 1290–1300.

A.R. Price, I.I. Voutchkov, G.E. Pound, N.R. Edwards, T.M. Lenton, S.J. Cox and the GENIE team, 2006. Multiobjective tuning of Grid-enabled Earth System Models using a Non-dominated Sorting Genetic Algorithm (NSGA-II) *Proceedings of the 2nd International Conference on eScience and Grid Computing*, Amsterdam, Netherlands.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3, `http://www.R-project.org/`.

J.C. Rougier, 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264.

J.C. Rougier, 2008. Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, **17**(4), 827–843.

J.C Rougier and D.M.H. Sexton, 2007. Inference in ensemble experiments. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2133–2143.

J.C. Rougier, D.M.H. Sexton, J.M. Murphy, and D. Stainforth, 2009. Analysing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *Journal of Climate*, **22(13)**, 3540-3557, DOI:10.1175/2008JCLI2533.1.

T.J. Santner, B.J. Williams, and W.I. Notz, 2003. *The Design and Analysis of Computer Experiments.* New York: Springer.

A. Schmittner, M. Latif and B. Schneider, 2005. Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations. *Geophyiscal research letters*, **32**, Art. No. L23710.

R.J. Stouffer, J. Yin, J.M. Gregory, K.W. Dixon, M.J. Spelman, W. Hurlin, A.J. Weaver, M. Eby, G.M. Flato, H. Hasumi, A. Hu, J.H. Jungclaus, I.V. Kamenkovich, A. Levermann, M. Montoya, S. Murakami, S. Nawrath, A. Oka, W.R. Peltier, D.Y. Robitaille, A. Sokolov, G. Vettoretti and S.L. Weber, 2006. Investigating the causes of the response of the thermohaline circulation to past and future climate changes. *Journal of Climate*, **19**, 1365-1387.

M. Vellinga and R.A. Wood, 2002. Global climatic impacts of a collapse of the Atlantic thermohaline circulation. *Climatic Change*, **54**, 251–267.

M. Vellinga and R.A. Wood, 2008. Impacts of thermohaline circulation shutdown in the twenty-first century. *Climatic Change*, **91**, 43–63.

W. N. Venables and B.D. Ripley, 2002. *Modern Applied Statistics with S.* New York: Springer-Verlag, fourth edition.

D. Würtz, 2007. High dimensional scrambled sobol sequences: An R and Splus software implementation. Unpublished, available at `http://www.itp.phys.ethz.ch/econophysics/R/pdf/JSS-Sobol.pdf`, and as part of the `fOptions` package.

K. Zickfeld, A. Levermann, M.M. Granger, T. Kuhlbrodt, S. Rahmstorff and D.W. Keith, 2007. Expert judgements on the response of the Atlantic meridional overturning circulation to climate change. *Climatic Change*, **82**, 235–265.