# Reified Bayesian Modelling and Inference for Physical Systems

Michael Goldstein, Jonathan Rougier *,[1]

*Department of Mathematical Sciences, University of Durham, U.K.*

**Abstract**

We describe an approach, termed *reified analysis*, for linking the behaviour of mathematical models with inferences about the physical systems which the models represent. We describe the logical basis for the approach, based on coherent assessment of the implications of deficiencies in the mathematical model. We show how the statistical analysis may be carried out by specifying stochastic relationships between the model that we have, improved versions of the model that we might construct, and the system itself. We illustrate our approach with an example concerning the potential shut-down of the Thermohaline Circulation in the Atlantic Ocean.

*Key words:* Computer Simulator, Emulator, Model Inadequacy, Structural Error, Thermohaline Circulation (THC), Model Design

## 1 Introduction: Using models to study physical systems

Many physical problems are studied through the construction and analysis of mathematical models. For example, our views about long term global climate change are largely governed by the analysis of large scale computer simulations for climate behaviour (see, e.g., Houghton *et al.*, 2001, ch. 8). While such analysis may be very revealing, it is also inherently limited, as even a good model will only offer an imperfect representation of the underlying physical system. This discrepancy raises fundamental questions as to how we should

* Corresponding author: Science Laboratories, South Road, Durham DH1 3LE, U.K.; tel +44(0)191 334 3111; fax +44(0)191 334 3051.

*Email addresses:* Michael.Goldstein@durham.ac.uk (Michael Goldstein), J.C.Rougier@durham.ac.uk (Jonathan Rougier).

[1] Funded by the U.K. Natural Environment Research Council (NERC), and the Tyndall Centre for Climate Change Research.

learn about actual physical systems through the analysis of models of the system. These questions go to the heart of the philosophy and practice of science (see, e.g., Cartwright, 1983; van Fraassen, 1989).

In this paper we treat our representation of the system as a function $f(x)$. It is helpful to distinguish between a model, its treatment and its simulator, where it is the *simulator* that is the function $f$. Broadly, we may think of this simulator as arising from

$$\mathsf{Simulator} = \mathsf{Model} + \mathsf{Treatment} + \mathsf{Solver}.$$

The model tends to be the underlying mathematical equations, often written as a collection of differential equations, equations of state, and, when involving several sub-domains, coupling equations. The treatment typically concerns the initial and boundary conditions and forcing functions that make the model applicable to a particular time, place, and scenario. The treatment can also concern which properties of the model are taken as outputs; e.g., steady state, 'ergodic' averaging, or dynamic evolution subject to specified forcing. Finally, the solver requires decisions about discretisations, in particular the order of the approximation and spatial and temporal resolution. When formulating a coherent framework linking models and systems it is essential to acknowledge that there are many *simulators* for a given system, sharing common features due to having similar models, treatments and solvers. We cannot think about how our given simulator $f$ is informative about a system without also thinking about how our simulator links to other simulators of the same system.

In our simulator $f$ we may consider the input vector $x$ as representing those aspects of the model and treatment that must be quantified before the simulator—typically represented as a computer code—will evaluate, and the output vector $f(x)$ to represent various features of the consequent behaviour of the physical system. For example, if $f$ was a representation of an oil reservoir, then $x$ might include a detailed enumeration of the geology of the reservoir, involving quantities such as local permeability and porosity of the rock structure, location and magnitudes of fault lines in the reservoir, and so forth, while $f(x)$ might represent time series of oil and gas production and pressure profiles at each of the wells in the reservoir (see, e.g., Thomas, 1982; Craig *et al.*, 1997). We do not know the geology of the reservoir, but we do have relevant prior knowledge. Some of this knowledge may be derived from observations from the system itself: we denote the system as $y$, and our observations of it (which may be incomplete or imprecise) as $z$. For example, we might have partial records of historical pressure readings and oil and gas production for some of the wells. The function $f$ is often complicated and high-dimensional, and can take several hours to evaluate for a single choice of input $x$, so that we must view the value of $f(x)$ to be unknown at each $x$ apart from that relatively small subset of $x$ values, say $X \triangleq \{x^{(1)}, \ldots, x^{(n)}\}$, at which we choose to eval-

uate the simulator outputs to give $F \triangleq \{f(x^{(1)}), \ldots, f(x^{(n)})\}$. Taken together, $S \triangleq (F; X)$ is referred to as the *ensemble* of evaluations.

There are various statistical questions associated with such an approach. In particular, what do the observation vector $z$ and the ensemble $S$ tell us about $x^*$, the 'correct' value of $x$ for the physical system? This is often termed *calibration*. Further, what do the values $\{z, S\}$ tell us about the value of the system vector $y$? This is often termed *system prediction*. For example, for the reservoir problem we may want to learn about the geology of the reservoir, and forecast the future oil and gas production of the reservoir under various management schemes.

To answer such questions, we must develop a joint probabilistic description which links $x^*$, $y$ and $\{z, S\}$; only in the case where the model, treatment and solver are all perfect and the data plentiful and measured without error might we be able to dispense with such a description, and treat the problem as a deterministic inverse problem. There are two issues to consider. Firstly, we need to construct a probabilistic specification of our beliefs for the function $f$. This is often termed an *emulator* for the function. We may use this emulator to update our beliefs about $f$ when we observe the model evaluations $S$. While this may be technically challenging if the form of $f$ is complicated and high dimensional, the general principles of emulation are reasonably well-understood, and there are many good references for the construction of effective emulators (sometimes referred to as *surrogates*): see, e.g., Currin *et al.* (1991), Craig *et al.* (1997, 1998), Kennedy and O'Hagan (2001), and the review in Santner *et al.* (2003, notably chapters 2–4).

The second issue is the relationship between the simulator $f$ and the actual system behaviour $y$. This is far less well-understood and indeed is very often ignored in the analysis of the simulator evaluations. Thus, the failure of the simulator to reproduce historical data $z$ is simply attributed to measurement error in the data and to a general but unquantified recognition of the inherent limitations of the approach. For example, while the analysis from global climate simulators is used to urge major changes in policy, we do not know of any treatment which has made a serious attempt to quantify uncertainty for the discrepancy between the simulators and actual climate behaviour, although probabilities are now being calculated which should depend on a careful treatment of such discrepancies (see, e.g., Murphy *et al.*, 2004). An analysis of future climate that includes the phrase "... conditional on the model (or the simulator) being correct" is not acceptable given the enormity of the consequences of an incorrect analysis, and the costs involved in adapting to or mitigating the problems of rapid climate change. Rougier (2006) discusses statistical issues in model-based predictions for future climate.

This paper is concerned with the logical framework within which a proper

uncertainty analysis may be conducted. In section 2 we describe the current 'state-of-the-art' for representing the relation between the simulator and the underlying system, and show that this is often incoherent and incomplete. Section 3 describes the general form of our approach, which we term *reified analysis*, and section 4 describes an implementation using linked *emulators*. The following sections carry out a reified analysis for an example concerning prediction of the potential shutdown of the Thermohaline Circulation in the Atlantic Ocean. The example is presented in section 5, the statistical modelling in section 6 and the results in section 7. The paper finishes with a brief summary in section 8.

## 2 Coherence issues in the Bayesian treatment

Much of the work on incorporating the uncertainties arising from simulator inadequacy (also referred to as *structural error*) has been carried out in the context of the current Bayesian treatment of computer code analysis; see Kennedy and O'Hagan (2001), Craig *et al.* (2001) and Higdon *et al.* (2004). In this treatment, we express simulator inadequacy through a relation of the form

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

where $x^*$ is the best input, $\epsilon$ is a random vector, termed the *discrepancy*, which is taken to be probabilistically independent of both $x^*$ and the function $f$, which we denote by $\epsilon \perp\!\!\!\perp \{x^*, f\}$. In Goldstein and Rougier (2004), a simulator for which our judgements obey (1) is termed a 'direct simulator'. Sometimes, minor modifications are made to (1), for example introducing a scalar regression multiplier on $f(x^*)$, but such modifications do not affect our discussion below. The magnitude of the variances of the components of $\epsilon$ corresponds to beliefs about how well the simulator represents the corresponding components of the physical system, and the covariance structure of $\epsilon$ expresses our beliefs about systematic errors across the simulator outputs. For example, in the hydrocarbon reservoir example, if the simulator substantially under-predicts pressure at a particular well for a sequence of time periods, is it likely that the simulator will similarly under-predict pressure for the next time period? In simulators where the output components are spatially and/or temporally indexed, we would expect the off-diagonal elements of $\mathsf{Var}(\epsilon)$ to reflect this structure: Craig *et al.* (2001, sec. 6.1) give an example.

This approach recognises and quantifies simulator inadequacy. It has no obvious competitors within a well-specified probabilistic framework, although the pressing need for calibrated simulators has spawned 'probability-like' approaches in particular application areas, e.g. Generalised Likelihood Uncertainty Estimation (GLUE) and its extensions in Hydrology (Beven and Binley, 1992; Beven, 2005). However, it does run into difficulties. Observe that (1)

implies that all of the information about $y$ that is contained in knowledge of the value of $x^*$ and the function $f$ may be summarised by the single function evaluation $f(x^*)$. This is a questionable assumption on two counts. Firstly, why should we judge—for our imperfect simulator—that there is *any* single best value in the simulator input space $\mathcal{X}$ for which we consider $f(x^*)$ to be sufficient for $x^*$ and $f$ in predicting $y$? Secondly, if there were such a single value $x^*$, why should that value be the value that we would measure in the system independently of the simulator? The simulator is usually constructed by simplifying the physics of the system (both in the model and the treatment) and approximating the solutions of the resulting equations. Therefore, there may only be an indirect relationship between the simulator inputs and the corresponding aspects of the physical system.

In practice, modellers often seem to take two somewhat contradictory positions about the status of the simulator's best input, on the one hand arguing that it is a hypothetical construct and on the other hand using knowledge and intuition derived from the physical system to set plausible intervals within which such a value should lie. Ocean simulators provide a well-documented example of this, where it is necessary to distinguish between *molecular viscosity*, the correct input value according to the underlying physical model, and *eddy viscosity*, the 'best' input value which can be several orders of magnitude larger (see, e.g., National Research Council (NRC), 1994, p. 171). However, it is hard to see why we should have any confidence in the ability of the simulator to forecast future system behaviour unless there is some relation between the values of the inputs to the simulator and the actual physical values for the system.

Such difficulties lead us to consider whether using (1) to link our particular simulator and the system may often be inconsistent with our wider beliefs. In particular, we will often be able to envisage the thought experiment of constructing an improved simulator, $f'$ say, which respects the behaviour of the physical system more closely than does $f$, for example by using the same model and treatment, but taking more care in the numerical solution of the underlying equations on finer discretisations in space and time. If we accept (1) for $f$, then we should presumably accept a similar relation for the better simulator $f'$, possibly with a different choice of best input value, $x^{*\prime}$ say. Therefore, in this case we would *simultaneously* have

$$y = f(x^*) + \epsilon \qquad \epsilon \perp\!\!\!\perp \{x^*, f\}\,, \tag{2a}$$

$$\text{and} \quad y = f'(x^{*\prime}) + \epsilon' \qquad \epsilon' \perp\!\!\!\perp \{x^{*\prime}, f'\}\,. \tag{2b}$$

To illustrate, consider the following simple case. We have a function with a scalar input and output, and the true value $x^*$ has a well-defined physical meaning. Suppose that, for the true function, $f'$, we do accept relation (1), i.e. $y = f'(x^*) + \epsilon'$, with $\epsilon' \perp\!\!\!\perp \{f', x^*\}$. However, suppose that it is expensive to

evaluate $f'$ for any $x$, so we produce a fast approximate solver, giving an approximate function $f$. Suppose that, from analysing many similar simulators, we consider that our beliefs about $f$ may be represented as

$$f(x) = b\,x + u(x) \tag{3}$$

where $b$ is an unknown scalar, and $u(x)$ is some stochastic process independent of $b$. The question at issue is whether, and when, we may also write $y = f(x^*) + \epsilon$, with $\epsilon \perp\!\!\!\perp \{f, x^*\}$? This property follows when our judgement is that, at an arbitrarily chosen $x$, the quantity $\Delta(x) \triangleq f'(x) - f(x)$ is independent of $\{u(\cdot), b\}$ and stationary, in which case our model for the true function would be

$$f'(x) = b\,x + u'(x) \tag{4}$$

where $u'(x) = u(x) + \Delta(x)$. We might make such a judgement if we consider that the regression structure that we expect to observe in $f$ is partially an artifact of the simplifications induced in the fast solver for the equations. However, if we consider that the regression structure uncovered in (3) is informative for the regression structure in $f'$, then it may be more natural to replace (4) with

$$f'(x) = b'x + u'(x) \tag{5}$$

where $b$ and $b'$ are correlated and $u'(x)$ is a stationary process independent of $b'$. With this specification, knowledge of $b$ and $x^*$ is informative for $f'(x^*) - f(x^*)$, so that property (1) for $f'$ is no longer consistent with the corresponding property for $f(x^*)$. In summary, it is natural firstly to impose requirement (1) on the more accurate function, and then to make a scientific judgement about the relation between the two functions to determine whether the requirement is also appropriate for the approximate function. In the above example, if our judgement supports relation (4), then what we gain from considering $f'$ is a natural way to decompose the variance of $\epsilon$ into two independent components, one component representing inaccuracies arising from simplifications in the solver, and the other representing simplifications in the representation of the system by $f'$. However, if our judgement supports relation (5), then, in addition, introducing $f'$ allows us to give a much more precise description of the relation between evaluations of $f$, the value of $x^*$, and the system value than we could otherwise meaningfully construct.

The general analysis for a pair of simulators follows similarly. We can express precisely the notion that $f'$ is better than $f$ in the form of the sufficiency condition

$$f' \succeq f \quad \Longleftrightarrow \quad y \perp\!\!\!\perp \{x^*, f\} \mid \{x^{*\prime}, f'\} \tag{6}$$

where '$\succeq$' denotes 'is at least as good as'. In other words, if we knew $\{x^{*\prime}, f'\}$ then $\{x^*, f\}$ would provide no additional information about $y$. Eq. (2) and condition (6) imply that $\epsilon' \perp\!\!\!\perp \{x^*, f\}$. From (2), we may write $\epsilon' \equiv f(x^*) -$

$f'(x^{*\prime}) + \epsilon$, and it follows that (2) and (6) imply that

$$f'(x^{*\prime}) - f(x^*) \perp\!\!\!\perp \{x^*, f\}\,. \tag{7}$$

But this implication is often counter-intuitive in practice. In many applications, we would expect that knowledge of $x^*$ would be informative for the value of $x^{*\prime}$ and that knowledge of $f$ would be informative for the form of $f'$, so that knowledge of $x^*$ and $f$ would be informative for $f'(x^{*\prime})$, and so for $f'(x^{*\prime}) - f(x^*)$.

Such considerations show that, far from being independent of $f$ and $x^*$, the discrepancy vector $\epsilon$ in (1) for our actual simulator very often has a complicated joint distribution with these quantities, so that in such cases we cannot consider that $\epsilon \perp\!\!\!\perp \{x^*, f\}$. Therefore we are required to specify a joint distribution for $\{\epsilon, x^*, f\}$, or else to admit a restrictive and often counter-intuitive form for our beliefs about our simulator with respect to an improved version.

In the above illustration, $f'$ is any possible better simulator. In important practical problems we will often have access to several actual simulators for the system. For example, there are a wide range of climate simulators which we may consult in forming views on long term climate behaviour, with different levels of accuracy and overlapping but not identical input spaces which arise from sharing some aspects of the underlying models and treatments. If we wish to use a representation of form (1) to integrate the information from each simulator analysis, then we must specify a version of (1) for each simulator, and then attempt to construct a joint belief specification over the discrepancy terms for each simulator. However, if we have misrepresented our beliefs for each individual simulator, then we have no obvious way to synthesise the collection of different simulator evaluations.

Therefore it is necessary that we formulate a coherent framework for our beliefs which accurately express the relationships which it is reasonable for us to hold when seeking to reconcile evaluations from a collection of simulators with beliefs about the physical system. In Goldstein and Rougier (2004) we introduced an alternative view of the relation between simulators and physical systems based on extending the formulation described above. We shall term this alternative approach *reified analysis*. To 'reify' is to consider an abstract concept to be real. In our context, therefore, a reified analysis is the statistical framework which allows us to move our inferences from the abstract notion of the mathematical model to the real notion of the physical system, via one or more simulators. In what follows, we develop a general approach for reified analysis, and then illustrate this approach with an example in ocean modelling.

## 3 General form of the reified analysis

We now describe our approach for linking one or more simulators with the underlying physical system. The sections that follow provide an illustration of our approach.

### *3.1 The reified simulator*

Suppose that we have a single simulator $f$. We might make the judgement that $f$ could be treated as a direct simulator, for which relation (1) would be appropriate. However, suppose we can envisage an improved simulator $f'$. If we replace relation (1) with (2b), then we may now consider a further improved version of $f'$, $f''$ say, and repeat this thought experiment. Therefore, to link the simulator and the physical system, we need to consider a version of the simulator with enlarged set of inputs, $f^*(x, w)$ say, which is sufficiently careful in respecting the physics of the system and which solves the resulting system of equations to a sufficiently high order of accuracy that we would not consider it necessary to make judgements about any further improvement to that simulator, in the following sense. If we consider a further improvement, $f^{**}(x, w, v)$ say, then we have no physical or mathematical insights as to how $f^{**}(x, w, v)$ may differ from $f^*(x, w)$, so that $f^{**}(x, w, v) - f^*(x, w)$ is independent of $\{f^*, x^*, w^*, v^*\}$.

Due to the high accuracy of $f^*$, it is credible to attribute to $f^*$ the corresponding property to (1), namely

$$y = f^*(x^*, w^*) + \epsilon^* \qquad \epsilon^* \perp\!\!\!\perp \{f, f^*, x^*, w^*\} \tag{8}$$

where $x^*$ and $w^*$ are *system* values, and $\epsilon^*$ is the residual discrepancy between our improved simulator and the system. This replaces assertions such as (1), which concern our actual simulator. We judge $\epsilon^*$ to be independent of $f^*$ and $\{x^*, w^*\}$, as we have removed the cause of our beliefs about any such dependency. We term $f^*$ the *reified simulator* for $f$. In most cases, we do not expect to be able to construct $f^*$ and evaluate it. Rather $f^*$ is a construct that permits us to offer a coherent and tractable account as to how evaluations of $f$, and partial knowledge of the physical quantities corresponding to the simulator inputs, may be combined to give a meaningful probabilistic description of the behaviour of the physical system. The ensemble $S$ (i.e., the evaluations of our actual simulator) is informative for $f$ and so for $f^*$. Partial knowledge about $\{x^*, w^*\}$ translates through (8) into partial knowledge about the value $y$ for the physical system. We summarise this as follows.

**Reifying Principle:** The reified simulator separates our actual simulator from the underlying system, as expressed in (8). Our actual simulator is informative for the underlying system because it is informative for the reified simulator.

### *3.2 Discussion of the Reifying Principle*

We formulated the reifying principle because of our experiences in trying to link computer simulator analysis with the performance of physical systems. The leading tractable Bayesian approach for relating the simulator and the system was to assume some version of the direct simulator property expressed by (1). However, in many cases, we saw no compelling scientific reason to adopt this property. Further, in our discussions with statistically numerate system experts, we found a similar reluctance to consider their uncertainties to be well represented by the direct simulator formulation. In particular, there was strong disagreement with an implication of (1), namely that there exists an input $x^*$ such that, were it to be known, only a single evaluation of the simulator would be necessary.

It seems natural to consider how different our analysis would be if we were to improve our simulator in various ways. Of course, it is a matter of judgement as to the level of detail to which this further elaboration should be taken. Further, we are reluctant to abandon completely the direct simulator property, as, particularly for large computer simulators with many inputs and outputs, some version of this assumption is important for the tractability of the Bayes analysis. Therefore, our suggestion is that the modeller should assess those features which are judged most important in improving simulator performance, both by more realistic modelling and by more careful solution methods. This leads us to consider a simulator, which we term the reified simulator, for which there are no substantial improvements that we can currently envisage of a kind such that we would currently wish to impose additional structure on the difference between the outputs of the two simulators. Therefore, we will be able to accept the direct property for the reified version of the simulator, without introducing obvious conflicts between the model and our underlying beliefs.

Sometimes, it will be straightforward to define operationally the form of the reified simulator; for example, we might be able to reify our simulator simply by improving its solver for the same underlying mathematical model. In other cases, because of the complexity of the system, we may not wish to fully detail the reified form. In such cases, we may identify the most important features of the improved simulator, and then add some extra variation to account for the remaining differences between the simulators; we follow this path in the

example that we analyse below. In all cases, the attempt to make a genuine representation of our uncertainties about the system will be of value, although the more carefully we consider the effects of simulator deficiencies, then the more benefit we will gain in specifying meaningful beliefs relating the simulator to the system.

We consider the reifying principle to be a sensible pragmatic compromise, which retains the essential tractability in linking our computer evaluations with the true system values to generate beliefs about system behaviour, removes certain obvious difficulties in other approaches by recognising and incorporating uncertainties arising from perceived deficiencies in the simulator, and provides a clear rationale for the joint modelling of the results of the analysis of several related simulators for the system. However, we are not dogmatic in advocating this principle. If modellers wish to use their simulator to make substantive statements about the underlying system, then it is incumbent on them to provide a clear rationale for the probabilistic judgements that they are using to link these, in principle, quite different things. We would be very interested to see any alternative methods for building a meaningful framework within which these linking judgements are naturally coherent, and which lead to tractable Bayesian analyses even for large systems.

Whichever approach we adopt, we should be prepared to invest a serious effort into the analysis of simulator discrepancy. For example, climate simulators can take years to model and program, the data used to calibrate the models is costly and time-consuming to obtain, and each evaluation of the simulator may cost many thousands of pounds, and take months. The climate community must therefore consider whether all this activity and expense is an end in itself, or whether the ultimate intention of all this effort is to make statements about the future behaviour of actual climate systems. Certainly, climate experts convey the impression that their models are informative for actual climate outcomes. However, if the intention is to make a realistic assessment as to how uncertain we should be about future climate behaviour, then this requires an effort of similar magnitude to each other aspect of the analysis, both to recognise and model all of the sources of uncertainty separating our simulators from the system, and also to quantify the magnitude of each such uncertainty. The reifying principle offers a more complete general description of the relevant uncertainties than any other formulation of which we are aware. However, in any particular application, the principle may oversimplify certain aspects of the discrepancy modelling, and we look forward to approaches that go beyond our formulation without sacrificing essential tractability.

Refied modelling raises interesting methodological questions. In particular, it is possible—at least partially—to validate our uncertainty assessment for any given simulator. However, there is no prospect of even a partial validation of the uncertainty assessment for the additional simulators that we introduce to

link our actual simulator with the system. In this sense, the additional simulators should be viewed as useful mental constructs which help us to relate our actual simulator evaluations to the physical system. The role of these constructs is to build meaningful joint beliefs between the system properties, the system performance and the simulator evaluations. Therefore, the relevant validation is to assess whether this joint specification is supported by our observation of the historical behaviour of the system. Different formulations for the implementation of the reified analysis result in different joint probabilistic specifications over the observables, and these may be compared by a variety of standard Bayesian approaches; see Goldstein (1991) for a particular method for comparing alternative belief specifications which is appropriate for competing mean, variance and covariance specifications.

Our reified approach has implications for certain standard types of analysis. For example, if the motivation for calibration is to learn about actual system values, or to predict future system performance, then, as our approach provides a joint description of beliefs about all of the relevant quantities, a Bayesian analysis provides a natural replacement for traditional calibration analyses, which is typically based on finding a single 'best fit' value for the simulator inputs. In a reified analysis it is not clear what it would mean to calibrate our original simulator. Suppose that our intention is simply to find a setting for the inputs for which the simulator outputs match reasonably closely to historical data and for which we have a reasonable degree of confidence in the predictions made at this input choice for future system behaviour. As we have a full description of uncertainty, we may therefore solve the optimisation problem of identifying the input value which minimises the difference between the simulator output and the system behaviour in some appropriate *probabilistic* metric. This provides a form of model calibration with an assessment of fit quality.

## 4 Reified analysis using emulators

### *4.1 The simplest case*

Consider the case of a single actual simulator $f$ and corresponding reified form $f^*$. We describe our beliefs about $f$ in terms of an emulator, represented, for a second-order analysis, as a mean function $\mathsf{E}\left(f(x)\right)$ for every $x$, and a covariance function $\mathsf{Cov}\left(f(x), f(x')\right)$ for every $(x, x')$ pair. These two functions may be specified directly, or they may be inferred from a more primitive specification. A common form for constructing the emulator using the latter approach is to combine both global and local effects: the $i^{\text{th}}$ component of $f(x)$, a scalar

value denoted $f_i(x)$, is expressed as

$$f_i(x) = \sum_j \beta_{ij}\, g_j(x) + u_i(x) \tag{9}$$

(see, e.g., Santner *et al.*, 2003, section 2.3). In this relation, the $g_j(\cdot)$ are known functions of $x$, where for simplicity we have chosen the same regressor functions for all $i$. The regression coefficients $\mathcal{C} \triangleq \{\beta_{ij}\}$ are unknown coefficients, so that the summation term represents those aspects of our beliefs about $f_i(x)$ which can be expressed through a simple global regression on appropriately-chosen functions of the inputs. The quantity $u_i(x)$, which we term the *emulator residual*, is usually taken to be *a priori* independent of the $\beta_{ij}$ values, and expresses beliefs about local variation away from the global regression at point $x$. The residual $u_i(\cdot)$ is usually taken to be either a weakly stationary process or a stationary Gaussian process, depending whether we intend to carry out a second-order analysis or a full probabilistic analysis; note that we do not take $u(\cdot)$ to be stationary in our illustration below, for reasons outlined in section 6.1. We must assign a covariance function for the vector process $u(\cdot)$.

If the dimension of the input vector $x$ is very high, then this approach to emulator construction is most successful when, for each output component $i$, there are a relatively small number of regressors which account for much of the variability of $f_i(x)$ over the range of $x$ values. These functions can be identified by a combination of expert elicitation, based on understanding of the physics of the system, and a careful choice of evaluations of the simulator, which allows us to assess the most important of these effects. Design, screening, and elicitation issues are discussed in, e.g., McKay *et al.* (1979), Sacks *et al.* (1989), Koehler and Owen (1996), Craig *et al.* (1998), and O'Hagan *et al.* (2006); the general principles of Bayesian experimental design are also relevant (see, e.g., Chaloner and Verdinelli, 1995).

Having constructed an emulator for $f$, one possible approach to reification is to consider how this emulator might be modified to emulate $f^*$. Suppose that we judge that there are qualitative similarities between $f$ and $f^*$—if we did not consider $f$ and $f^*$ to have any such similarities, then it is hard to see what information about the physical system could be obtained by evaluating $f$. A natural way to express this similarity would be to express our emulator for $f^*$ as

$$f_i^*(x, w) = \sum_j \beta_{ij}^*\, g_j(x) + \sum_j \theta_{ij}^*\, h_j(x, w) + u_i^*(x, w); \tag{10}$$

where we write the collection of regression coefficients in the reified emulator as $\mathcal{C}^* \triangleq \{\{\beta_{ij}^*\} \cup \{\theta_{ij}^*\}\}$. The summation term in the $\theta_{ij}^*$ is introduced to account for additional systematic variation in the enlarged input space. Unless we have specific views to the contrary, we may construct this term to represent a source of variation uncorrelated with the variation accounted for in $f$ so that the components of $h$ are orthogonal to those of $g$, according to

our prior specification for $\{x^*, w^*\}$, and the $\theta^*_{ij}$ quantities are likewise uncorrelated with the other coefficients. Therefore, the relationship between $f$ and $f^*$ is contained in the relationship between the $\beta_{ij}$ and the $\beta^*_{ij}$ coefficients for each emulator, and between $u(x)$ and $u^*(x, w)$. Examples of how we might treat these relationships are given in our illustration in sections 6.2 and 6.3.

The overall effect of this type of approach is to impose an appropriate structure on our beliefs about the discrepancy between the actual simulator $f$ and the behaviour of the system, $y$. This discrepancy is divided into (i) a highly structured part reflecting our beliefs about the effects of more detailed considerations designed to improve the accuracy of $f$, and (ii) a possibly less structured part reflecting our views as to the absolute limitations of the type of simulator represented by $f$.

### *4.2 Inference using the reified analysis*

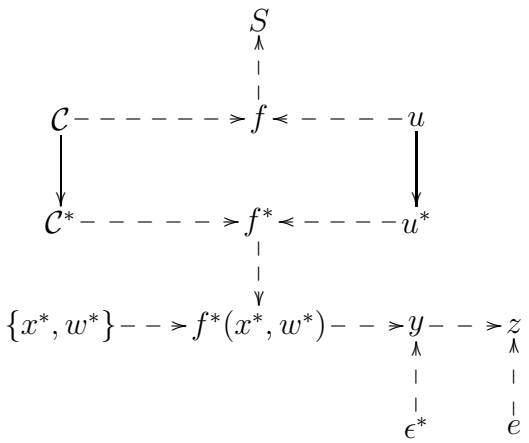
Given the above specification, our inferential calculations are, in principle, straightforward. The ensemble $S$ is informative for the coefficients $\mathcal{C}$ and the residual function $u(\cdot)$ in the emulator for $f$, (9). This information modifies our beliefs about $\mathcal{C}^*$ and $u^*(\cdot)$, and thus modifies beliefs about $f^*$ through (10). These modified beliefs are combined with our assessment of $\{x^*, w^*\}$ and $\epsilon^*$ to derive beliefs about the system value $y$, using (8). Deriving beliefs about $y$ in this way (i.e. by propagating uncertainty through a simulator, possibly via an emulator) is known as *uncertainty analysis* (see, e.g., Haylock and O'Hagan, 1996; O'Hagan *et al.*, 1999).

We may also want to incorporate information from observations made on the system, denoted $z$. In general we can write these observations as $z = H(y) + e$ for some known function $H(\cdot)$, where $e$ denotes a vector dominated by observation errors, taken to have mean zero (for simplicity) and to be independent of all other uncertain quantities. For a general $H(\cdot)$ we would create new simulator outputs which we would then emulate directly. In the special case where $H(\cdot)$ is a linear transformation we can proceed with an emulator for $f^*(\cdot)$. This is appropriate for our illustration below, and so we express the transformation in terms of the incidence matrix $H$, writing

$$z = Hy + e \qquad H \text{ specified, } e \perp\!\!\!\perp \text{ all other quantities.} \tag{11}$$

Often we can arrange for $z$ and components of $y$ to correspond one-to-one, but this form also includes the case where the data and the system values do not match-up precisely, and interpolation or averaging is required. With (11), we have a joint belief specification for $\{y, z\}$, and we can update our beliefs about $y$ using the observed value for $z$, namely $\tilde{z}$.

We can represent the relationship between our data $\{z, S\}$ and our prediction $y$ in terms of a *Bayesian belief network (BBN)* (see, e.g., Cowell *et al.*, 1999):




(12)

'Child' vertices that are strictly determined by their 'parents' are indicated with dashed lines. Therefore the probabilistic modelling in this BBN involves specifying the independent marginal quantities $\mathcal{C}$, $u(\cdot)$, $\{x^*, w^*\}$, $\epsilon^*$ and $e$, and the conditional quantities $\mathcal{C}^* \mid \mathcal{C}$ and $u^*(\cdot) \mid u(\cdot)$.

For our inferences below, we adopt the approach of Craig *et al.* (2001). This approach is based on a second-order specification for the emulators and the link between the emulators, the system and the system observations. Craig *et al.* propose a two-stage approach. For the first stage, the mean and variance of $f^*(x^*, w^*)$ are computed by integrating $\{x^*, w^*\}$ out of the emulator for $f^*$. According to our reified approach, this emulator is constructed by building a joint emulator for $\{f, f^*\}$, then updating this joint emulator using the ensemble $S$, and then marginalising the result to infer the emulator for $f^*$. For the second stage, the mean and variance for $f^*(x^*, w^*)$ are used to construct the joint mean and variance for the collection $\{y, z\}$; then beliefs about $y$ are adjusted using the observed value $z = \tilde{z}$. A more detailed analysis of this approach is given in Goldstein and Rougier (2006).

The *Bayes linear* approach describes the appropriate adjustment for a second-order prior specification; see, e.g., Goldstein (1999). Denote the mean and variance of $f^*(x^*, w^*)$ as $\mu$ and $\Sigma$ respectively. Then, using (8) and (11), the Bayes linear adjusted mean and variance for $y$ given $z = \tilde{z}$ are

$$\mathsf{E}_{\tilde{z}}(y) = \mu + V_y H^T (H V_y H^T + V_e)^{-1} (\tilde{z} - H\mu) \tag{13a}$$

$$\mathsf{Var}_{\tilde{z}}(y) = V_y - V_y H^T (H V_y H^T + V_e)^{-1} H V_y \tag{13b}$$

where $V_y \triangleq \mathsf{Var}(y) \equiv \Sigma + \mathsf{Var}(\epsilon^*)$, and $V_e \triangleq \mathsf{Var}(e)$. In general we can make this calculation scale more-or-less costlessly in the number of simulator inputs and outputs, by choosing regressor functions in the emulators that are orthonormal with respect to a weighting function proportional to the prior

distribution of $\{x^*, w^*\}$; in this way the integration over $\{x^*, w^*\}$ reduces the regressor functions to 0s and 1s. A similar approach using orthonormal regressors is described in Oakley and O'Hagan (2004).While this scalability is crucial for large applications, for our illustration below we will use simple non-orthonormal regressor functions for clarity, and perform the integration numerically.

### *4.3 Structural reification*

Our formulation is intended to offer the maximum flexibility to the analyst to consider the structure of beliefs relating our actual simulator to its reified counterpart, and to the system. Much of this structure will derive from specific improvements which we might consider building into our simulator. Our beliefs about the impact of specific modifications, which we term *structural reification*, form an important part of our specification relating the emulators for $f$ and $f^*$, as they directly reflect on our beliefs about the strengths and weaknesses of $f$ as a representation of the physical system.

One natural thought experiment for the modeller is to consider $f$ to be embedded in a larger simulator $f'$, in the sense that inputs to $f'$ are $(x, v)$, where there is some set of values in $v$, $\mathcal{V}_0$ say, for which $f'$ replicates $f$, so that, for each $x$,

$$f(x) = f'(x, v_0) \qquad v_0 \in \mathcal{V}_0. \tag{14}$$

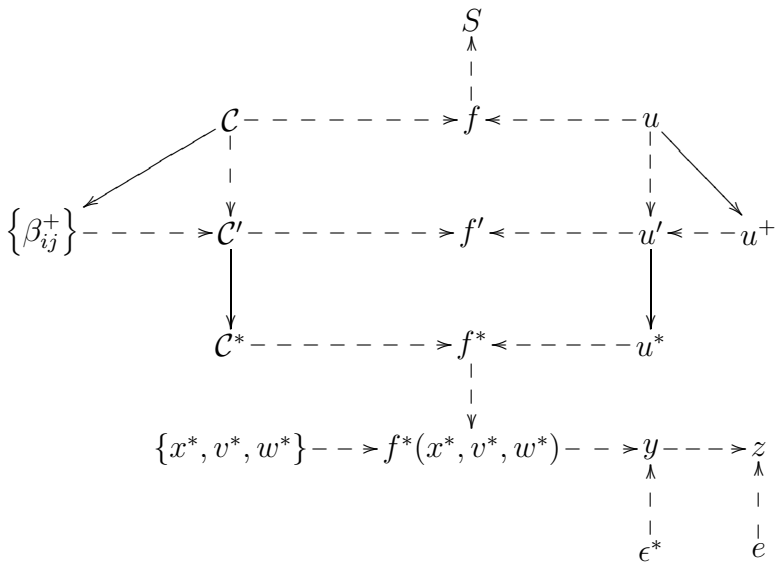
We call $f'$ with this property a *generalisation* of $f$; it represents the simplest form of structural reification. If our emulator for $f$ is built as (9), then a simple way to build the emulator for $f'$ while respecting constraint (14) is to create an emulator for $f'$ 'on top' of that for $f$ :

$$f'_i(x, v) = f_i(x) + \sum_j \beta^+_{ij}\, g^+_j(x, v) + u^+_i(x, v) \tag{15}$$

which has additional regressors $g^+(\cdot)$, and an additional residual $u^+(\cdot)$; necessarily $g^+_j(x, v_0) = 0$ and $u^+_i(x, v_0) = 0$ for $v_0 \in \mathcal{V}_0$. We write $\mathcal{C}' \triangleq \mathcal{C} \cup \{\beta^+_{ij}\}$ for the combined set of coefficients in the generalised emulator. The additional terms in (15) cause the emulator for $f'$ to deviate from that for $f$ when $v \notin \mathcal{V}_0$. We may choose to treat these additional terms independently of $\mathcal{C}$ and $u(\cdot)$, but it seems more natural to link them together. We will illustrate this in section 6.2.

Now we consider how we might join $f$, $f'$ and $f^*$ in the case where $f'$ is a generalisation of $f$. In this case we must ensure that the joint structure of the three emulators satisfies the markov property that $f'$ separates $f$ and $f^*$, because $f'$ encapsulates $f$ in its entirety. A simple way to enforce the markov structure across our emulators is to have $\mathcal{C} \perp\!\!\!\perp \mathcal{C}^* \,|\, \mathcal{C}'$, and $u \perp\!\!\!\perp u^* \,|\, u'$.

We can illustrate the joint structure of the three emulators in a BBN which extends (12):




(16)

In the limit as $v \to \mathcal{V}_0$, $\mathcal{C}'$ tends to $\mathcal{C}$ and $u'(x, v)$ to $u(x)$; at which point we are back at (12).

The BBN in (16) is a template for the way in which information from the ensemble $S$ gets passed to the system value $y$ via the simulators, and in particular via the regression coefficients and the residual processes in the emulators. We control the passage of information according to our judgement of how similar the three simulators are. Specifically, our choices for the conditional distributions $\{\beta_{ij}^+\}|\mathcal{C}$ and $u^+(\cdot)|u(\cdot)$ quantify our judgement regarding the relationship between $f$ and $f'$, and our choices for $\mathcal{C}^* \,|\, \mathcal{C}'$ and $u^*(\cdot) \,|\, u'(\cdot)$ do the same for $f'$ and $f^*$. A natural metric for the size of the distances between the simulators is the mean variances of the differences

$$\begin{aligned} \Delta' &\triangleq \mathsf{E}\left(\mathsf{Var}\left(f'(x^*, v^*) - f(x^*) \mid x^*, v^*\right)\right) \\ \text{and } \Delta^* &\triangleq \mathsf{E}\left(\mathsf{Var}\left(f^*(x^*, v^*, w^*) - f'(x^*, v^*) \mid x^*, v^*, w^*\right)\right); \end{aligned} \tag{17}$$

these can be compared with the discrepancy variance, which gives us the distance between $f^*(x^*, v^*, w^*)$ and $y$:

$$\mathsf{Var}\left(\epsilon^*\right) = \mathsf{E}\left(\mathsf{Var}\left(y - f^*(x^*, v^*, w^*) \mid f^*, x^*, v^*, w^*\right)\right),$$

as $\epsilon^* \perp\!\!\!\perp \{f^*, x^*, v^*, w^*\}$ and $\mathsf{E}\left(\epsilon^*\right) = \mathbf{0}$.

The reified approach is flexible enough to be extended to the situation in which we have more than one actual simulator; e.g., where we have two ensembles $S$ and $S'$, where $S'$ may be from the simulator $f'$ described in the previous subsection, or $S'$ may be from another simulator entirely, sharing only limited aspects of the underlying model and treatment. For example, in climate prediction we can predict system features such as climate sensitivity to a doubling of atmospheric $CO_2$ using recent observations on the climate state vector, or using palæo-climate data collected from sources such as sedimentary pollen, fossilised trees and ice-cores. We must combine these two sources of information coherently in a manner that takes account of the features that the palæo and contemporary simulators share.

At a formal level, the extension to two or more ensembles from different simulators requires us to join extra vertices like $S'$ to the appropriate quantities in the belief net describing the joint relationship between the simulators. We envisage the reified simulator to be accurate enough that we judge it to be sufficient for the collection of actual simulators, so that $\epsilon^*$ in (8) can be treated as independent of that collection. The practical details involved in this joint specification and analysis may be challenging, particularly if $f'$ does not separate $f$ and $f^*$, but we know of no other general, tractable approach for meaningfully unifying the analyses derived from a collection of simulators with overlapping input spaces.

### *4.5 Model design*

Our structured reification involves a particular generalised simulator $f'$. As we have constructed the joint belief specification over the collection $(f, f', f^*)$ we may evaluate the benefit that we should expect from actually constructing and evaluating $f'$. We call this assessment *model design*. Model design has a similar role to that of experimental design. In experimental design, we evaluate the potential for various experiments to reduce our uncertainties for unknown quantities of interest. Such analyses are invaluable in helping us to make efficient use of our limited resources, and in particular in warning us against undertaking experiments which are highly unlikely to achieve our objectives.

Model design has a similar purpose. Constructing and evaluating a large-scale simulator is an extremely time-consuming activity. Therefore, it is very important to develop analytical tools which will guide our judgements concerning the ability of such simulators to reduce our uncertainty about important features of the underlying system. However, while there is an enormous literature about

experimental design, we know of no systematic methodology which serves a comparable function for the construction of simulators of physical systems. Structural reification does offer such a methodology, which we illustrate in the context of our example in section 7.3.

## 5 Example: Thermohaline Circulation in the Atlantic

In this section we describe a physical system, the inference that we wish to make about it, a simulator that is informative for that inference, and the structural reification that we use to link our actual simulator and the system.

### *5.1 The Thermohaline Circulation*

The system is the Thermohaline Circulation (THC) in the Atlantic Ocean. The THC is the mechanism by which heat is drawn up from the tropics towards the western seaboard of Europe. At the moment there is concern about the effect of global warming on the THC, because changing temperature and precipitation patterns will alter the temperature and salinity characteristics of the Atlantic. The extreme case is THC shutdown, which could significantly lower the temperature of the western seaboard of Europe. One important quantity in the Atlantic is the amount of freshwater re-distribution that would cause this shutdown to occur. This is the quantity that we wish to predict, and to assist us we have data on other aspects of the Atlantic, namely its temperature and salinity, and the current size of the THC.

### *5.2 The basic model*

We base our analysis on the recent paper by Zickfeld *et al.* (2004), hereafter 'ZSR'. The ZSR model of the Atlantic is a four-compartment Ordinary Differential Equation (ODE) system, shown schematically in Figure 1. Each compartment is described by its volume and its depth. The state vector comprises a temperature, $T_i(t)$, and salinity, $S_i(t)$, for each of the four compartments. Freshwater re-distribution is modelled by the two parameters $F_1$ and $F_2$, and atmospheric temperature forcing by the three parameters $T_1^*$, $T_2^*$ and $T_3^*$.

The key quantity in the model is the rate of meridional overturning, $m(t)$, which is the flow-rate of water through the compartments and proxies the THC. Overturning is assumed to be driven linearly by temperature and salinity

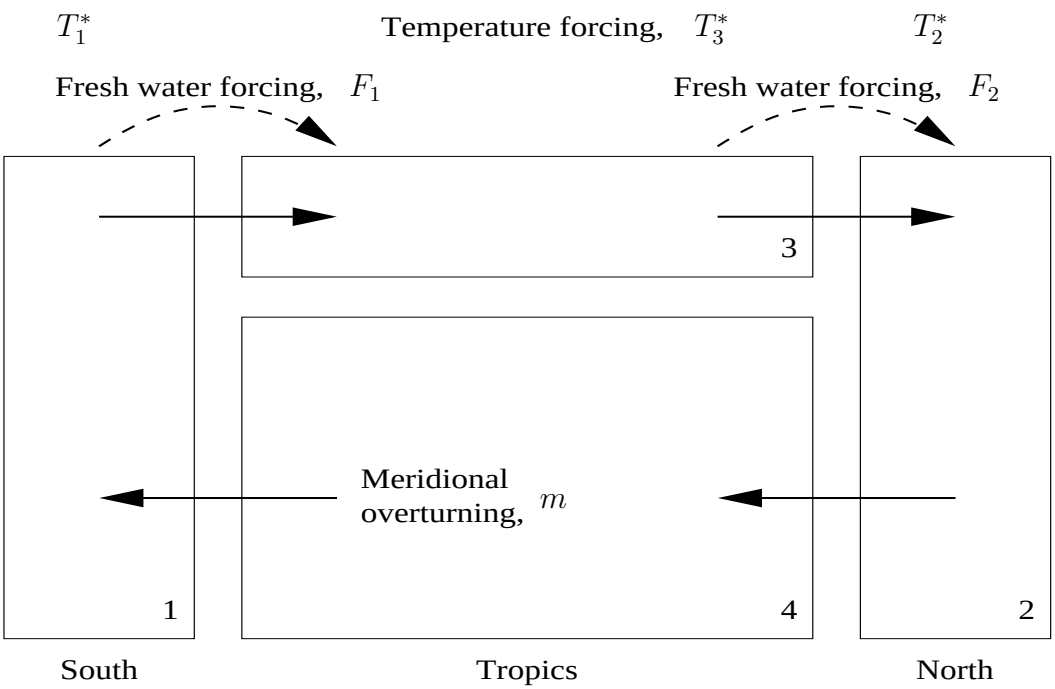



Fig. 1. A compartmental model of the Atlantic, as described in Zickfeld *et al.* (2004). Arrows indicate the direction of positive flow.

differences between compartments 1 and 2,

$$m(t) = K\left\{\beta[S_2(t) - S_1(t)] - \alpha[T_2(t) - T_1(t)]\right\} \tag{18}$$

where $\alpha$ and $\beta$ are thermal and haline expansion factors, and $K$ is an empirical flow constant. Hence $m(t)$ tends to be bigger at times when the 'northern Atlantic' is colder and more salty than the 'southern Atlantic'. Large values of $F_1$ make the 'south Atlantic' more salty, and tend to reduce $m(t)$.

### *5.3 Our treatment of the model*

We can treat the ZSR model in a number of different ways. Here we consider an aspect of its equilibrium properties, where by 'equilibrium' we mean the steady state with all model parameters specified and time-invariant. For a given value of the parameters we can compute the equilibrium value of the state vector, and the equilibrium value for $m$; these equilibrium values are indicated with an 'eq' superscript. We can also find the value of $F_1$ at which equilibrium overturning collapses to zero, denoted $F_1^{\rm crit}$; this is clarified after eq. (21). The value of $F_1^{\rm crit}$ is of particular interest because the current value of $F_1$ is thought to be close to $F_1^{\rm crit}$. If it is under $F_1^{\rm crit}$ then the impact of an increasing $F_1$ on the THC is reversible. But if it goes beyond $F_1^{\rm crit}$ then the THC, once collapsed, cannot easily be restarted by reducing $F_1$ (as shown

in ZSR, Figure 2). This *hysteresis* in the THC was noted by Stommel (1961), and is one of the mechanisms by which gradual changes in climatic forcing, due to global warming, might bring about abrupt and, on centennial scales, irreversible changes in regional climate.

This treatment maps the parameters into 8 quantities (we will not need the temperature and salinity values for compartment 4). We follow ZSR in fixing all but five of the parameters, treating only $T_1^*$, $T_2^*$, $T_3^*$, $\Gamma$ and $K$ as uncertain (the $\Gamma$ parameter appears in the ODE system). Furthermore, because we believe that the three temperatures satisfy the ordering $T_2^* \leq T_1^* \leq T_3^*$, we reparameterise these as

$$\tau_1 \triangleq T_2^*, \quad \tau_2 \triangleq T_1^* - T_2^*, \quad \tau_3 \triangleq T_3^* - T_1^* \tag{19}$$

as this ordering will be part of our beliefs about the relationship between our model and the system, described in section 6.4. This gives us a vector of simulator inputs $x$, where

$$x \triangleq (\tau_1, \tau_2, \tau_3, \Gamma, K); \tag{20}$$

to help with interpretation, we map each of the inputs into the range $[0, 1]$, using the ranges for the components of $x^*$ given in Table 1. We then define our simulator $f$ as

$$f_i(x) \triangleq \begin{cases} T_i^{\mathrm{eq}}(x) & i = 1, 2, 3 \\ \Delta S_{21}^{\mathrm{eq}} \triangleq S_2^{\mathrm{eq}}(x) - S_1^{\mathrm{eq}}(x) & i = 4 \\ \Delta S_{32}^{\mathrm{eq}} \triangleq S_3^{\mathrm{eq}}(x) - S_2^{\mathrm{eq}}(x) & i = 5 \\ m^{\mathrm{eq}}(x) & i = 6 \\ F_1^{\mathrm{crit}}(x) & i = 7 \end{cases} \tag{21}$$

and the superscript 'eq' denotes denotes the equilibrium value when $F_1 = 0.014$, thought to be its current value. Note that $F_1^{\mathrm{crit}}$ is a different type of output from the others. If we think of overturning as the function $m(x, F_1)$, i.e. with $F_1$ as an extra input, then $m^{\mathrm{eq}}(x) \triangleq m(x, 0.014)$, and $F_1^{\mathrm{crit}}(x)$ satisfies $m\Big(x, F_1^{\mathrm{crit}}(x)\Big) \equiv 0$. Note also that (18) gives us an explicit form for $f_6(x)$ in terms of $x$ and three of the other outputs; this was one of our motivations for modelling salinity differences rather than salinity levels. We could exploit this knowledge to construct a better joint emulator for $f(x)$, and this would be an important part of a detailed treatment of the problem; for this illustration, however, we choose to treat all of the outputs in the same manner, to simplify the analysis.

Table 1
Simulator inputs, definitions, units and ranges.

| | Definition | Units | Interval for $(x^*, v^*)$ | |
|---|---|---|---|---|
| | | | Lower | Upper |
| | *Inputs in the original simulator* | | | |
| $x_1$ | $\tau_1 \triangleq T_2^*$ | °C | 0 | 10 |
| $x_2$ | $\tau_2 \triangleq T_1^* - T_2^*$ | °C | 0 | 5 |
| $x_3$ | $\tau_3 \triangleq T_3^* - T_1^*$ | °C | 0 | 10 |
| $x_4$ | $\Gamma$ | $\mathrm{W\,m^{-2}\,°C^{-1}}$ | 10 | 70 |
| $x_5$ | $K$ | Sv | 5,000 | 100,000 |
| | *New inputs in the generalised simulator* | | | |
| $v_1$ | $q$ | (none) | 0 | 0.3 |
| $v_2$ | $T_5^*$ | °C | 0 | 10 |

### *5.4 Structural enhancements to the simulator*

There are two obvious ways in which we might improve our simulator. First, we might allow additional model parameters to be uncertain, giving a simulator with a larger input space. Second, we might generalise the ZSR model itself. Often we will pursue both of these routes, but to avoid over-complicating our example we will illustrate only the second in this paper. This tends to correspond to the way that models often get generalised in practice. For example, large climate models have sub-models for processes such as cloud and sea-ice formation, glaciation and ocean/atmosphere coupling, and contain switches to introduce or exclude certain features. The Supplementary Information accompanying Murphy *et al.* (2004) provides an example of such switches in a large model from the Hadley Centre.

For our generalisation of the model, we choose to add on an additional compartment at the southern end, representing the other oceans. We model the THC as though some of it bleeds off into these other oceans, rather than circulating only within the Atlantic. The generalised model is shown in Figure 2, with the extra compartment numbered 5, and compartment 1 split vertically into $1A$ and $1B$. A fixed proportion $q$ of $m$ is bled off from compartment $1B$ into compartment 5. The same volume of water returns from compartment 5 into compartment $1A$, but it carries with it the temperature and salinity characteristics of the other oceans. We treat these as fixed, with uncertain

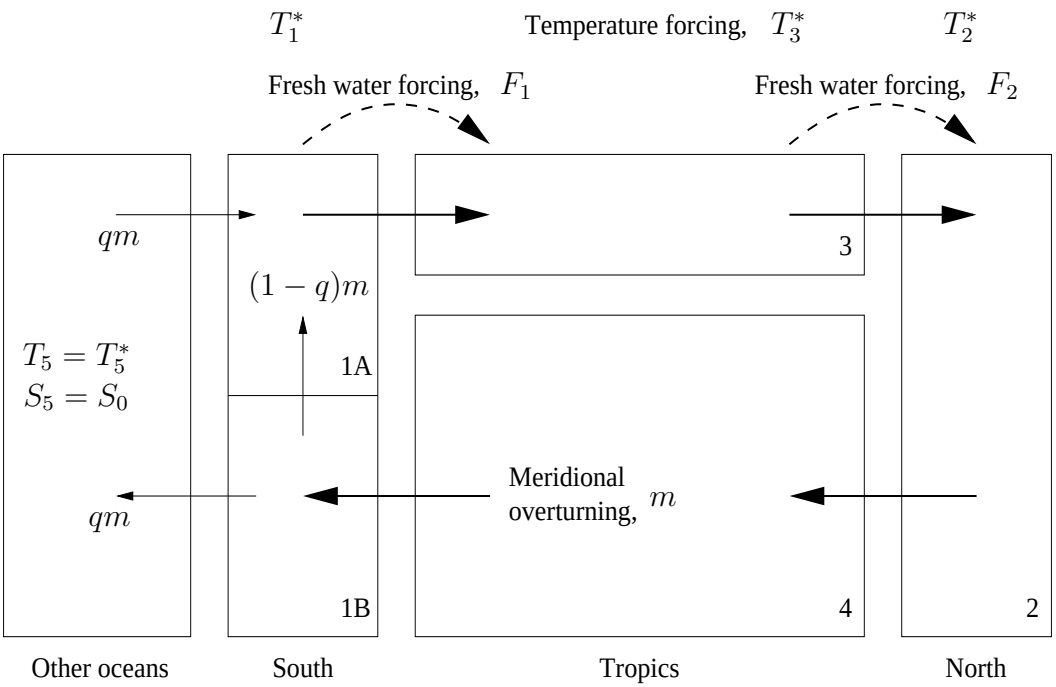



Fig. 2. A generalisation of the compartmental model of Zickfeld *et al.* (2004).

temperature $T_5^*$ and given salinity $S_0$, the reference salinity. Once we define

$$T_1 \triangleq \frac{V_{1A}}{V_{1A}+V_{1B}}\,T_{1A} + \frac{V_{1B}}{V_{1A}+V_{1B}}\,T_{1B} \tag{22}$$

and likewise for $S_1$, eq. (18) still applies, and the interpretation of the output of the generalised simulator is the same as that of the original. Our generalised simulator is $f(x,v)$ where $v \equiv (v_1, v_2) \triangleq (q, T_5^*)$. We recover the original simulator with $q = 0$, regardless of the value of $T_5^*$, i.e.

$$f'(x, 0, v_2) = f(x) \quad \text{for all } x \text{ and } v_2. \tag{23}$$

In terms of the outline in section 4.3, $\mathcal{V}_0 = \{v : v_1 = 0\}$.

## 6 Example (cont): Emulators and reification

This section describes the statistical framework in which we specify a joint mean and variance for the collection of emulators $\{f, f', f^*\}$, incorporating information from the ensemble $S$. Our general strategy in this illustration is to start with $(f : S)$, which denotes an emulator for $f$ constructed from the ensemble, then to specify $f'$ conditionally on $(f : S)$, and then to specify $f^*$ conditionally on $f'$. This sequential approach to constructing our reified emulator has much in common with that of Reese *et al.* (2004), who propose

a sequential method for integrating expert judgement, simulator evaluations and experimental data in a Bayesian hierarchical model. In its implementation, however, Reese *et al.*'s approach makes exactly the assertion that we are seeking to avoid, by assimilating expert judgement, the *actual* simulator evaluations and the system data, all into the same object. For the reasons described in section 2, we are concerned both that this assertion may be inappropriate in many applications, and that in applying it we may materially oversimplify the correlation structure.

We choose to start the quantification of our uncertainty about the function with $(f : S)$ in this illustration because this allows us to avoid formulating our prior for $f$; this is useful in situations where we have limited information about $f$ beyond that contained in the ensemble $S$. The alternative—which is more appealing both foundationally and when the ensemble $S$ is small relative to the dimension of the input space—is to specify a proper prior for the collection $\{f, f', f^*\}$, and then update this whole collection with $S$. Our decision to start our treatment with $(f : S)$ is a pragmatic compromise in situations where it is hard to specify an appropriate prior for $f$, either because $f$ is poorly understood, or where there is sufficient information in the ensemble $S$ that the effort involved in a careful specification of the prior is unlikely to be worthwhile. Therefore this section discusses constructing the emulator $(f : S)$, section 6.1, constructing the emulator $f'$ using the emulator $(f : S)$, section 6.2, constructing the emulator $f^*$ using the emulator $f'$, section 6.3, and specifying $y$ in terms of $f^*(x^*, v^*)$, section 6.4. At the second and third stages our choices are guided by our judgements regarding the differences between the simulators, and, at the fourth stage, between the system and the reified simulator at its best inputs.

### *6.1 Emulating our actual simulator*

A general form for the emulator of our actual simulator was given in (9). For the regressors $g(\cdot)$ we use a constant and linear terms only. We consider the residual to comprise additional terms of the general form $u(x) = Ah(x)$ where $A$ is a matrix of unknown coefficients and the components of $h(\cdot)$ are specified functions that are orthogonal to $g(\cdot)$ with respect to a uniform weight function, in this case linear combinations of low-order monomial terms such as $(x_1)^2$, $x_1x_2$, and so on, as described in An and Owen (2001). We impose a matrix normal distribution on $A$ to simplify the structure of the residual, with the row variance being specified in terms of hyperparameters, and the column variance estimated along with the regression coefficients conditional on the hyperparameters. The hyperparameters themselves are fitted by cross-validation.

Our approach to emulator construction differs from the more standard approach (e.g., Currin *et al.*, 1991; Kennedy and O'Hagan, 2001), which simplifies the residual by asserting prior stationarity, and for which the hyperparameters control the correlation function of the residual directly, rather than implicitly. Our approach ensures that, for an orthogonal design, there is no correlation between our updated uncertainty for the regression coefficients $\mathcal{C}$ and our updated uncertainty about the residual $u(x)$. This is an important consideration in reified modelling, where each has a different role to play. Our approach also makes possible a more detailed treatment of the residual functions across emulators, for example by linking up coefficients in the residual coefficient matrices, $A$, $A'$ and so on; in our illustration, however, we simply treat the residuals in terms of their mean and variance functions. The strengths of the linear fits in most components (see Table 2) suggest that alternative approaches to emulating our simulator are unlikely to reach substantially different conclusions.

We build our emulator (9) for $f$ directly from an ensemble of just 30 evaluations in a maximin Latin Hypercube in the inputs, rather than by formally updating from a prior specification for $\mathcal{C}$ and $u(\cdot)$. This small number of evaluations reflects the practical limitations of many large simulators. The resulting coefficient means are given in Table 2 (these are the GLS fitted values), along with the $R^2$ values from the fit on the linear regression terms, and the square root of the averaged (over $x$) residual variance, remembering that the updated residual function is non-stationary. The three temperature outputs are strongly but not exactly determined by linear terms in the three temperature inputs; the two salinity differences are less well determined by the linear regressors (but better-determined than the individual salinities); the final two outputs, $m^{\mathrm{eq}}$ and $F_1^{\mathrm{crit}}$, are reasonably well determined. As a simple diagnostic, the leave-one-out prediction errors (after first fitting the emulator hyperparameters) are shown in Figure 3; the 7-vector of prediction errors at each $x$ has been standardised using the predictive mean vector and variance matrix to be uncorrelated with mean zero and variance one; experience suggests that these are a sensitive measure of an emulator's ability to summarise and extrapolate the ensemble. This diagnostic is border-line acceptable; we regard our emulator as reasonable for the purposes of illustration.

To simplify our emulator for $f$, we zero the updated correlations between the regression coefficients and the residual coefficients: these are non-zero only because the 30-point design in the model-inputs is not exactly orthogonal. Then the mean function of our emulator interpolates the points in the ensemble, but the variance function does not go to zero at these points.

Table 2
Emulator regression coefficients (means). The constant represents the unconditional expectation and the five inputs are scaled to have the same range. The $R^2$ value indicates the squared correlation between the actual values and the fitted values from the linear regression terms. The average residual standard deviation (ARSD) indicates the square root of the residual variance after uniformly averaging over the input space.

| | Const. | $\tau_1$ | $\tau_2$ | $\tau_3$ | $\Gamma$ | $K$ | $R^2$ (%) | ARSD |
|---|---|---|---|---|---|---|---|---|
| $T_1^{\mathrm{eq}}$ | 7.451 | 2.898 | 1.325 | 0.024 | −0.010 | −0.028 | ∼100 | 0.075 |
| $T_2^{\mathrm{eq}}$ | 5.693 | 2.898 | 0.518 | 0.117 | −0.154 | 0.297 | 99 | 0.221 |
| $T_3^{\mathrm{eq}}$ | 12.391 | 2.878 | 1.405 | 2.851 | 0.036 | −0.043 | ∼100 | 0.040 |
| $\Delta S_{21}^{\mathrm{eq}}$ | −0.154 | −0.008 | 0.095 | −0.050 | 0.037 | 0.124 | 60 | 0.173 |
| $\Delta S_{32}^{\mathrm{eq}}$ | 0.234 | −0.025 | −0.048 | 0.053 | −0.021 | −0.102 | 60 | 0.139 |
| $m^{\mathrm{eq}}$ | 10.218 | −0.357 | 10.396 | −0.920 | 3.453 | 6.337 | 90 | 3.627 |
| $F_1^{\mathrm{crit}}$ | 0.087 | 0.001 | 0.068 | −0.005 | 0.014 | 0.021 | 90 | 0.021 |

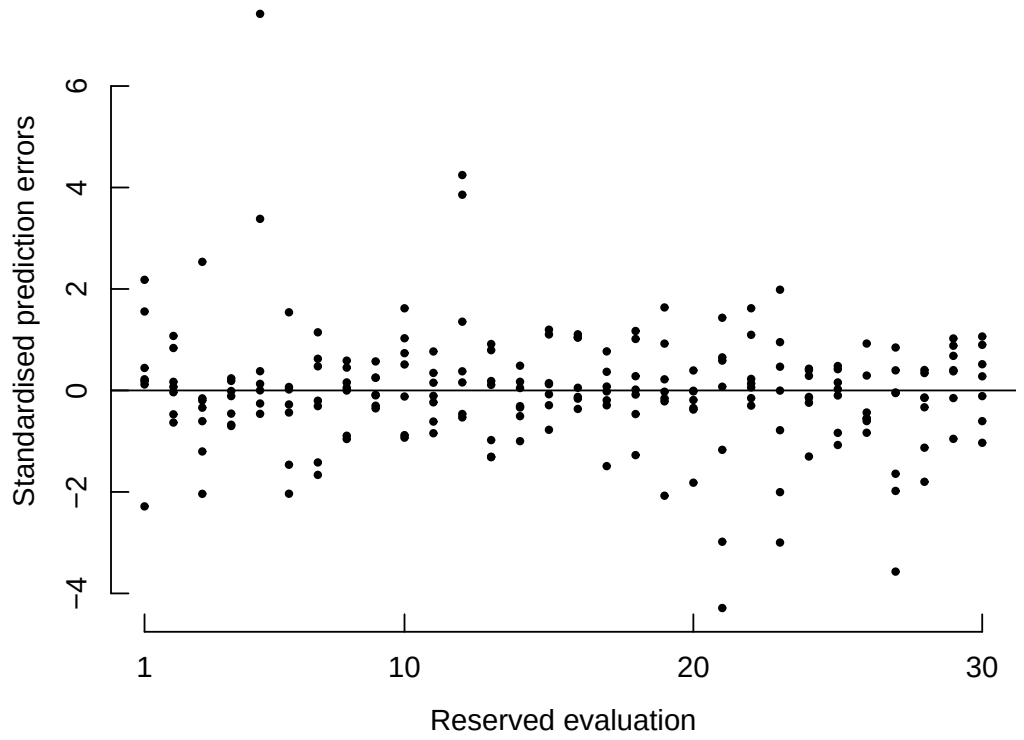



Fig. 3. Leave-one-out emulator diagnostic, showing standardised prediction errors for the seven outputs for each of the 30 evaluations in the ensemble.

In section 5.4 we considered constructing a simulator with an extra compartment, representing the other oceans, denoted $f'$. Our emulator for $f'$ must satisfy (23), so that it reduces to $f$ in the special case where the other oceans play no role. In our case $v_1 = 0$ (i.e. $q = 0$) is the value which collapses the generalised simulator back to the original. To simplify our choices below, we rescale $v_1$ and $v_2$ to the range $[0, 1]$ using the values given in Table 1.

Our judgement is that, overall, $f$ and $f'$ will be quite similar. More specifically, in terms of the original quantities, we judge that introducing the extra compartment with a sizable $q$ will tend to dampen the response of the model outputs to the relaxation temperatures $T_1^*$, $T_2^*$ and $T_3^*$. We also judge that a sizable $q$ will decrease $m^{\text{eq}}$ and $F_1^{\text{crit}}$, partly directly and partly through decreasing the temperature and increasing the salinity in compartment 1.

We now describe how we quantify these judgements. We stress that the choices we make are not arrived at in isolation, but also taking account of the resulting properties of the emulators themselves. As described in section 5.4, one way for us to assess our choices is in terms of balancing the distance between $f$ and $f'$, $f'$ and $f^*$, and $f^*$ and $y$. For our choices, described below, these distances are summarised in Table 3.

### *6.2.1 Structure of our generalised emulator*

We base our emulator for $f'$ on (15), introducing the additional regression terms

$$g_j^+(x, v) = \begin{cases} v_1 \times g_j(x) & j \in \{0, \ldots, 5\} \\ v_1 & j = 6 \\ v_1 \times v_2 & j = 7, \end{cases} \tag{24a}$$

and the additional residual term

$$u_i^+(x, v) = v_1 \times \Big\{ c_i \, u_i(x) + \delta_i^+(x, v) \Big\} \tag{24b}$$

for some specified value $c_i$ and an additional mean-zero random function $\delta_i^+(\cdot)$, for each output. These additional terms satisfy the property that $g_j^+(x, v_0) = u_i^+(x, v_0) = 0$ for $v_0 \in \mathcal{V}$. Then we can write

$$f_i'(x, v) \equiv \sum_{j=0}^{7} \beta_{ij}' \, g_j'(x, v) + u_i'(x, v) \tag{25a}$$

where we set $\beta_{ij} = 0$ for $j \in \{6, 7\}$, and

$$\beta'_{ij} \triangleq \beta_{ij} + v_1\, \beta^+_{ij}, \tag{25b}$$

$$g'_j(x, v) \triangleq \begin{cases} g_j(x) & j \in \{0, \ldots, 5\} \\ g^+_j(x, v)/v_1 & j \in \{6, 7\}\,, \end{cases} \tag{25c}$$

$$u'_i(x, v) \triangleq (1 + v_1\, c_i)\, u_i(x) + v_1\, \delta^+_i(x, v). \tag{25d}$$

In terms of our BBN representation of the relationship between emulators, given in (16), we have $\mathcal{C} \triangleq \{\beta_{ij}\}$ and $\mathcal{C}' \triangleq \{\beta'_{ij}\}$.

### *6.2.2 Matched coefficients in the two emulators*

For the matched coefficients, i.e. those for which $j \in \{0, \ldots, 5\}$, our judgement regarding how $\beta_{ij}$ and $\beta'_{ij}$ differ according to the value of $v_1$ translate into a specification for the mean and variance of $\{\beta^+_{ij}\} \mid \{\beta_{ij}\}$. To specify this conditional relation we use the general framework

$$\beta^+_{ij} = (c_{ij} + \omega_{ij})\, (\beta_{ij} - m_{ij}) + (r_i/r_j)\, \nu_{ij} \tag{26}$$

where $m_{ij}$ and $c_{ij}$ are given scalars, and $\omega_{ij}$ and $\nu_{ij}$ are independent mean-zero random quantities with specified variances. The scalars $r_i$ and $r_j$ denote typical scales for the relevant output and regressor, respectively; their role is to allow us to specify $\mathsf{Sd}\,(\nu_{ij})$ in scale-free terms ($\omega_{ij}$ is already scale-free). We will use ranges for $r_i$ and $r_j$, where $r_i$ is inferred from the ensemble, and $r_j$ from the $g'_j(\cdot)$.

Substituting (26) into the expression for $\beta'_{ij}$ and re-arranging gives

$$\beta'_{ij} - m_{ij} = \Big(1 + v_1\, (c_{ij} + \omega_{ij})\Big)\, \Big(\beta_{ij} - m_{ij}\Big) + v_1\, (r_i/r_j)\, \nu_{ij}\,. \tag{27}$$

The $m_{ij}$ represent offsets to ensure that the $\beta'_{ij}$ are appropriately centred. We set all of the $m_{ij}$ to zero for $j \in \{1, \ldots, 5\}$, but we use non-zero values for some of the constants, namely $m_{i0} = 8$ for $i \in \{1, 2, 3\}$, to centre the temperatures away from zero. The $c_{ij}$ can be used to shrink or expand $\beta'_{ij} - m_{ij}$ relative to $\beta_{ij} - m_{ij}$. We represent our specific judgements given above as

$$c_{ij} = \begin{cases} -0.10 & i \in \{1, \ldots, 7\} \text{ and } j \in \{1, 2, 3\} \\ -0.05 & i \in \{6, 7\} \text{ and } j = 0 \\ -0.05 & i = 1 \text{ and } j = 0 \\ -0.05 & i = 4 \text{ and } j = 0 \end{cases}$$

and 0 otherwise. Treating both $c_{ij}$ and $\nu_{ij}$ as small, the $\omega_{ij}$ describe the probability of a change of sign between $\beta_{ij} - m_{ij}$ and $\beta'_{ij} - m_{ij}$. We set $\mathsf{Sd}\,(\omega_{ij}) = 1/3$

for all matched coefficients, so that a reversal of sign when $v = 1$ is approximately a three-standard-deviation event, i.e. has less than 5% probability according to the three-sigma rule for unimodal marginal distributions (Pukelsheim, 1994). The $\nu_{ij}$ describe the variation in $\beta'_{ij}$ in the event that $\beta_{ij} = m_{ij}$ with probability one. We set $\mathsf{Sd}(\nu_{ij}) = 1/18$ for all matched coefficients, so that when $\beta_{ij} = m_{ij}$ the additional term for regressor $j$ will—with high probability—account for less than a sixth of the range of its response when $v_1 = 1$. Choosing small values for $\mathsf{Sd}(\omega_{ij})$ and $\mathsf{Sd}(\nu_{ij})$ is one way in which we express our judgement that the two simulators $f$ and $f'$ are similar.

### *6.2.3 New coefficients in the generalised emulator*

For the new coefficients in $\{\beta^+_{ij}\}$ for $j \in \{6, 7\}$ we use a similar but simpler framework, namely

$$\beta'_{ij} \equiv v_1 \, \beta^+_{ij} = v_1 \, (r_i / r_j) \, \nu_{ij} \tag{28}$$

where the $\nu_{ij}$ have the same properties as in (26). We set $\mathsf{Sd}(\nu_{ij}) = 1/9$, reflecting our view that each of the new terms in the emulator for $f'$ will—with high probability—account for less than one third of the range of its response when $v_1 = 1$.

### *6.2.4 The residual in the generalised emulator*

We believe that our generalisation will make the behaviour of the simulator a more complicated function of the inputs. Therefore we want the variance attributable to the residual at any given $(x, v)$ to be larger for $f'$ than for $f$, except in the limit as $v_1 \to 0$. We choose $c_i = -0.1$, and treat $\delta^+(\cdot)$ as an independent second-order stationary random process with a variance matrix equal to the expected variance of $u(\cdot)$. Effectively we are shrinking the structured component of the residual and then adding an unstructured process, so that we are partly 'decorrelating' $u(\cdot)$ and $u'(\cdot)$ by our specification for $u^+(\cdot)$. With these choices, the standard deviation of $u'_i(\cdot)$ is about 35% larger than that of $u_i(\cdot)$ when $v_1 = 1$.

Note that it is not necessary for us to specify a covariance function for $\delta^+(\cdot)$, nor for $\delta^*(\cdot)$ which occurs in section 6.3. Our inference requires that we can compute the mean and variance of the emulators pointwise (section 4.2); we do not require, for example, the covariance between $f'$ at two different input values. This would not have been the case had we started with a proper prior specification for $\{f, f', f^*\}$ and then conditioned on $S$. Our inference is insensitive to the covariance structure of $\delta^+(\cdot)$ and $\delta^*(\cdot)$ because we summarise the information in our ensemble in the emulator $(f : S)$, and because of the sequential way in which we treat $f'$ as a generalisation of $(f : S)$, and $f'$ as sufficient for $f^*$, as shown in (16).

Table 3
Unconditional mean and standard deviation for the actual simulator; square root of the distances between this simulator, the generalised simulator and the reified simulator, as defined in (17); unconditional mean and standard deviation for the reified simulator; and standard deviation of the discrepancy.

| | $f(x^*)$ | | $\sqrt{\Delta'}$ | $\sqrt{\Delta^*}$ | $f^*(x^*, v^*)$ | | $\mathsf{Sd}\,(\epsilon^*)$ |
|---|---|---|---|---|---|---|---|
| | Mean | Std dev. | | | Mean | Std dev. | |
| $T_1^{\mathrm{eq}}$ | 7.451 | 3.188 | 1.181 | 2.009 | 7.464 | 3.813 | 0.840 |
| $T_2^{\mathrm{eq}}$ | 5.693 | 2.982 | 1.103 | 2.138 | 5.691 | 3.706 | 0.840 |
| $T_3^{\mathrm{eq}}$ | 12.391 | 4.289 | 1.644 | 3.372 | 12.389 | 5.516 | 0.840 |
| $\Delta S_{21}^{\mathrm{eq}}$ | −0.154 | 0.282 | 0.134 | 0.236 | −0.151 | 0.351 | 0.075 |
| $\Delta S_{32}^{\mathrm{eq}}$ | 0.234 | 0.219 | 0.112 | 0.213 | 0.235 | 0.296 | 0.075 |
| $m^{\mathrm{eq}}$ | 10.218 | 13.540 | 5.150 | 9.916 | 9.940 | 16.907 | 3.300 |
| $F_1^{\mathrm{crit}}$ | 0.087 | 0.079 | 0.033 | 0.066 | 0.085 | 0.103 | 0.044 |

### *6.3 Emulating the reified simulator*

We now emulate the reified simulator. On the basis of our previous choices we implement (10) as the reification of (15), so that

$$f_i^*(x, v) = \sum_{j=0}^{7} \beta_{ij}^* \, g_j'(x) + u_i^*(x, v) \tag{29}$$

where to simplify our account we have not introduced any further simulator inputs or regressor functions. In this formulation each $\beta_{ij}^*$ relates directly to $\beta_{ij}'$, where the $\{\beta_{ij}'\}$ were described in section 6.2; in our BBN, given in (16), $\mathcal{C}^* \triangleq \{\beta_{ij}^*\}$.

Our judgement is that the distance between $f'$ and $f^*$ is larger than that between $f$ and $f'$, i.e. typically $\Delta_{ii}^* > \Delta_{ii}'$ using the definitions in (17). This reflects our view that $f$ and $f'$ are quite similar, but that further extensions to the simulator, for example the additional of further new compartments or subdivisions of existing compartments, could have a larger impact. With extra compartments, spatially coarse inputs such as the three relaxation temperatures could be disaggregated. We have not done this, and therefore some of our uncertainty about how $f^*$ responds to the inputs follows from the fact that by retaining these inputs at their coarse resolution, we may be introducing non-linear aggregation effects.

We summarise this judgement in our choices for $\{\beta^*_{ij}\} \mid \{\beta'_{ij}\}$ and $u^* \mid u'$. For the regression coefficients our implementation is

$$\beta^*_{ij} - m_{ij} = (1 + \omega^*_{ij})\,(\beta'_{ij} - m_{ij}) + (r_i/r_j)\,\nu^*_{ij} \qquad j \in \{0, \ldots, 7\} \tag{30}$$

where the $m_{ij}$ play the same role as in (26) and $\omega^*$ and $\nu^*$ have similar properties to before. We set $\mathsf{Sd}\left(\omega^*_{ij}\right) = 1/2$ for all $i$ and $j$, so that a sign-change between $\beta'_{ij} - m_{ij}$ and $\beta^*_{ij} - m_{ij}$ is a two-standard-deviation event, i.e. less unlikely than a sign change between $\beta'_{ij} - m_{ij}$ and $\beta_{ij} - m_{ij}$, and we set $\mathsf{Sd}\left(\nu^*_{ij}\right) = 1/18$ for all $i$ and $j$. For the residual, our implementation is

$$u^*_i(x, v) = c^*_i\, u'_i(x, v) + \delta^*_i(x, v) \tag{31}$$

where we choose $c^*_i = 3/4$ for all $i$, and $\delta^*(\cdot)$ is an independent second-order stationary process with variance equal to the average variance of $u(\cdot)$.

As in section 6.2, these choices for the components of the emulator for $f^*$ are not made in isolation, but with careful attention paid to their consequences for the behaviour of the emulator. Table 3 shows the resulting distance measure $\Delta^*$, and the unconditional mean and variance of $f^*(x^*, v^*)$. By comparing the two columns $\sqrt{\Delta'}$ and $\sqrt{\Delta^*}$ we can see that, through the modelling choices we have made, the distance between $f^*$ and $f'$ is roughly twice the distance between $f'$ and $f$. We consider this reasonable as $f^*$ does not represent the 'perfect' simulator, but simply a simulator accurate enough that we are prepared to assert relation (8) for $f^*$ and $y$.

### *6.4 System values*

The system in our case is the Atlantic. However, the very high level of aggregation in our simulator makes it more appropriate to use as data the output of a larger climate simulator, which has been carefully tuned to the Atlantic in a separate set of experiments. We can think of the larger simulator as a component of a highly sophisticated measuring device that is used to quantify aspects of the Atlantic. The further role of an advanced model as a component of a measuring device raises some interesting issues, which we shall not explore here, as our intention in this illustration is to stay in broad agreement with the analysis in ZSR, who calibrate their model to data from the CLIMBER-2 intermediate complexity coupled ocean/atmosphere simulator. This simulator provides values for the three equilibrium temperatures $T_1^{\mathrm{eq}} = 6$, $T_2^{\mathrm{eq}} = 4.7$ and $T_3^{\mathrm{eq}} = 11.4$, the salinity differences $S_2^{\mathrm{eq}} - S_1^{\mathrm{eq}} = -0.15$ and $S_3^{\mathrm{eq}} - S_2^{\mathrm{eq}} = 0.25$, and the equilibrium overturning $m^{\mathrm{eq}} = 22.6$ (ZSR, Table 3).

From the reified approach, our statistical framework linking the reified simu-

lator, the system and the system observations is

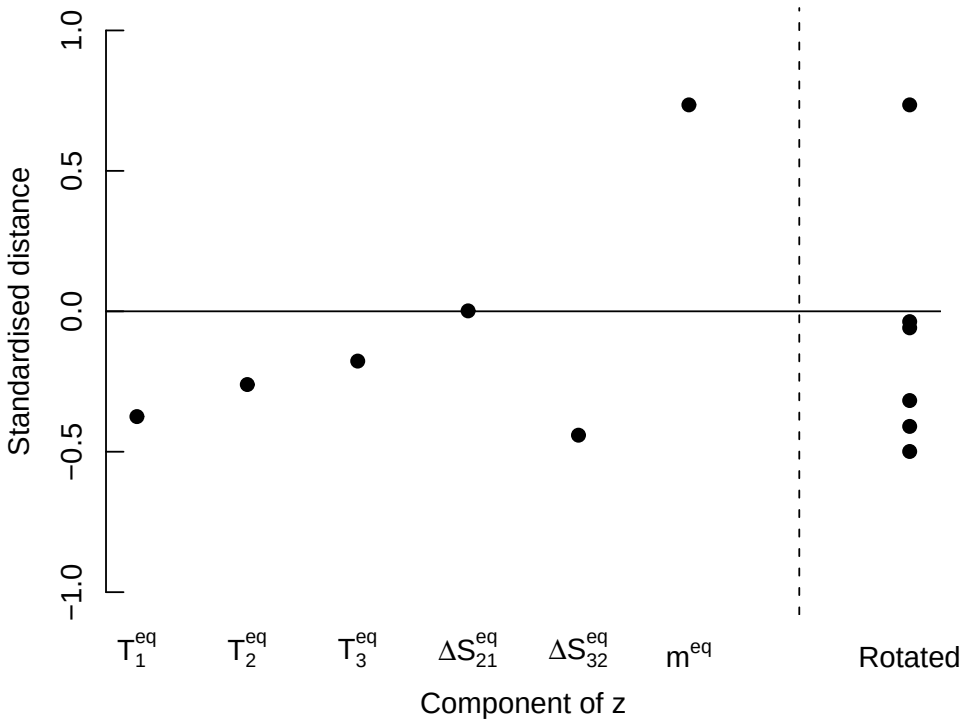
$$y = f^*(x^*, v^*) + \epsilon^* \qquad \text{and} \qquad z = Hy \equiv \Big(y_1, \ldots, y_6\Big) \tag{32}$$

where $y$ denotes the system values corresponding to our simulators' outputs, and $z$ denotes the observations that we collect from CLIMBER-2, which comprise the first six components of $y$, as described immediately above. To be in agreement with ZSR, we adopt a mean and variance for $(x^*, v^*)$ that is consistent with independent uniform distributions for each of the five components, based on the ranges given in Table 1.

The discrepancy $\epsilon^*$ is independent of $\{f, f', f^*, x^*, v^*\}$. We set its expectation to zero, and for the variance we choose a mostly-diagonal matrix with individual component standard deviations given in the final column of Table 3. These values reflect our judgement concerning the relationship of the three simulators and the system. Our starting point was that the distance between $y$ and $f^*$, as summarised by $\mathsf{Var}\,(\epsilon^*)$, would be smaller but the same order of magnitude as the distance between $f^*$ and $f'$, as summarised by $\Delta^*$. For the six equilibrium values we specified a standard deviation for $\epsilon_i^*$ that was one-third of $\sqrt{\Delta_{ii}^*}$, based on the mean value over outputs of the same type. For the seventh output, $F_1^{\text{crit}}$, which is more complicated, we used two-thirds. All standard deviations were expressed to two significant digits. We also included a correlation of $-0.5$ between the two salinity differences in $\epsilon^*$ (components four and five) to account for the shared term, salinity in compartment 2.

**Model validation.** It is not possible to validate the emulators $f'$ and $f^*$ directly. However, the role of these emulators is to lead us to a more appropriate joint distribution between all of the actual observables, namely the ensemble of simulator evaluations, the system observations, and the system itself. Therefore, a natural diagnostic is to compute the predictive mean and variance of $z$, and compare these to the observed value $\tilde{z}$. Figure 4 shows the standardised marginal prediction errors for each of the components of $z$, as well as the collection of all six values after transforming to uncorrelated quantities with mean zero and variance one. The standardised distances are small, but overall we feel they offer a reasonable validation of our statistical choices, and we prefer to leave these as they stand rather than risk 'over-tuning'. In a more critical analysis we would look again at the various sources of uncertainty about $z$, and see if we could identify an obvious candidate for reduction.

Fig. 4. Prediction errors for the data, $z = \tilde{z}$. The first six columns show the standardised errors for each component, and the final column shows the collection of errors transformed to be uncorrelated with mean zero and variance one.




## 7 Example (cont): Reified inference for $F_1^{\text{crit}}$

### *7.1 Main results*

For our predictions we are primarily interested in $y_7 \equiv F_1^{\text{crit}}$, summarised by the adjusted mean and variance $\mathsf{E}_{\tilde{z}}(y_7)$ and $\mathsf{Var}_{\tilde{z}}(y_7)$, as given in (13). The amount by which we can reduce our uncertainty about $F_1^{\text{crit}}$ will depend on the correlations between the components of $z$ and $y_7$. For example, on the basis of our ensemble $S$ and our statistical modelling choices, $m^{\text{eq}}$ and $F_1^{\text{crit}}$ have a correlation of 0.46, and so we expect that adjusting by $z = \tilde{z}$, which includes an observation on $m^{\text{eq}}$, will improve our prediction for $F_1^{\text{crit}}$.

The current value for $F_1$ is thought to be about 0.014 Sv, as measured using CLIMBER-2 (ZSR, Table 1). It is of fundamental interest to determine how close this value is to $F_1^{\text{crit}}$, the value at which the THC will shut down. For this illustration we report a mean and standard deviation for $F_1^{\text{crit}}$, and the standardised distance between the current value of $F_1$ and the mean of $F_1^{\text{crit}}$. Our prediction is summarised in Table 4. On our prior assessment, $F_1$ is about 0.6 standard deviations below the mean critical value. After adjusting by $z = \tilde{z}$,

Table 4
Predictions of $F_1^{\text{crit}}$. The current value of $F_1$ is thought to be about 0.014 Sv; the final column gives the standardised distance between this value of $F_1$ and $\mu$, the mean of $F_1^{\text{crit}}$.

| | Mean, $\mu$ | Std dev., $\sigma$ | $(F_1 - \mu)/\sigma$ |
|---|---|---|---|
| Initial, based on 30 evaluations | 0.085 | 0.112 | −0.631 |
| After adjusting by $z = \tilde{z}$ | 0.119 | 0.099 | −1.061 |

the mean of $F_1^{\text{crit}}$ rises and its standard deviation falls; our adjusted assessment shows that $F_1$ is about 1.1 standard deviations below the mean critical value.

### *7.2 Sensitivity analysis*

The choices we made in section 6 were based on our subjective beliefs about the relation between $f$, $f'$, $f^*$ and $y$. However, while these beliefs seem plausible order-of-magnitude representations, they are not the result of a careful expert scientific analysis. We investigate our choices in a simple experiment over: first, $c_{ij}$, $\mathsf{Sd}\,(\omega_{ij})$, $\mathsf{Sd}\,(\nu_{ij})$, $c_i$ and the parameters $\mathsf{Sd}\,(\delta_i(\cdot))$, which describe the relationship between $f$ and $f'$; second, $\mathsf{Sd}\left(\omega^*_{ij}\right)$, $\mathsf{Sd}\left(\nu^*_{ij}\right)$, $c^*_i$ and $\mathsf{Sd}\left(\delta^*_i(\cdot)\right)$, which describe the relationship between $f'$ and $f^*$; and third, $\mathsf{Sd}\,(\epsilon^*)$, which describes the relationship between $f^*(x^*, v^*)$ and $y$. For each of these quantities we try halving and doubling our choice, and we present the results in terms of the adjusted mean and standard deviation of $F_1^{\text{crit}}$, and the standardised distance between the current value of $F_1$ and $F_1^{\text{crit}}$.

The results are shown in Table 5. This confirms that the distance between $f$ and $f^*$ directly affects our uncertainty about $F_1^{\text{crit}}$, and about the relationship between $F_1$ and $F_1^{\text{crit}}$: larger standard deviations on quantities such as $\omega_{ij}$ and $\omega^*_{ij}$ increase the distance between $f$ and $f^*$, and introduce more uncertainty about $f^*$ when starting from our ensemble of evaluations of $f$. Overall, however, our assessment of the event $F_1 < F_1^{\text{crit}}$ seems to be quite robust to our choices for these parameters, with the standardised distance typically lying between −0.9 and −1.2 standard deviations. The two extremes values (−0.62 and −1.39) come from varying $\mathsf{Sd}\left(\omega^*_{ij}\right)$, which is the most influential parameter according to Table 5.

We now illustrate the issue of model design, as discussed in section 4.5. We compare two options: doing more evaluations on $f$, or constructing $f'$ and doing evaluations on that instead. We can investigate the benefit of the first option by supposing that we could do sufficient evaluations on $f$ to reduce our uncertainty about the regression coefficients and residual to zero. We can approximate this state by constraining $\mathcal{C}$ and $u(\cdot)$ in the emulator for $(f : S)$ to their expected values, i.e., by zeroing their variances (which are in general not zero). With this treatment the adjusted standard deviation for $F_1^{\text{crit}}$ is 0.097, only a 2% reduction in uncertainty. For our choices in the emulator, the 30 evaluations that we already have are highly informative about $f$.

The best possible case in the second option would be to construct $f'$ and do sufficient evaluations to reduce our uncertainty about $\mathcal{C}'$ and $u'(\cdot)$ to zero. Following the same procedure, the adjusted standard deviation for $F_1^{\text{crit}}$ is 0.089, about a 10% reduction in the uncertainty. This is much larger than in the first option, and suggests that constructing and the evaluating the simulator $f'$ will provide substantial information that is not available in evaluations of $f$. In line with our findings for $f$, we would expect that most of the benefit from building $f'$ would come from the early evaluations, in a carefully-chosen design.

## 8 Conclusion

In this paper, we have described an approach, which we term *reified analysis*, for linking the behaviour of simulators based on mathematical models with inferences about the physical systems which the simulators represent. We have two motivations for such an approach. Firstly, it is of fundamental importance to clarify the logical basis for making assertions about physical systems given the analysis of simulators which are known to be far from perfect. Reified analysis offers a far more meaningful treatment of the uncertainty relating the model analysis and the behaviour of the physical system than does any other approach that we are aware of, and addresses fundamental concerns about the limitations of the current approach.

Secondly, reified analysis offers a structural approach for assessing all of the uncertainties which arise in relating collections of simulators to the system that those simulators represent. In our illustration we have shown some simple ways to structure the reified analysis, so that choices about the relationships between the various simulators, both actual and conceptual, can be specified in an intuitive and relatively scale-free way. We believe that it is not only more flexible but also simpler to think of the relationship between our actual

simulator $f$ and the system $y$ in a series of steps rather than one big step. By representing these steps in terms of a small collection of parameters, we obtain a reasonable trade-off between flexibility and simplicity.

The relationship between the simulator and the system is, typically, a subtle and complicated matter, and the resulting statistical constructions and belief specification may be challenging. However, this merely emphasises the importance of carrying out such analysis carefully within a clear and coherent framework. Reified analysis offers such a framework and allows us to express our understanding of the strengths and weaknesses of our simulators to whatever degree of detail we find helpful.

## Acknowledgements

We would like to thank Kirsten Zickfeld for her help with the Zickfeld *et al.* (2004) model of the Atlantic; we stress that our treatment of this model, while attempting to be broadly comparable with that in Zickfeld *et al.*, is entirely our own. We would also like to thank Peter Challenor for identifying a problem with an earlier calculation, and the Editor and Referees for their very perceptive comments.

## References

J. An and A.B. Owen, 2001. Quasi-regression. *Journal of Computing*, **17**, 588–607.

K. Beven, 2005. A manifesto for the equifinality thesis. *Journal of Hydrology*. In press.

K. Beven and A. Binley, 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–298.

G.E.P. Box and G.C. Tiao, 1973. *Bayesian Inference in Statistical Analysis*. Reading, Massachusetts: Addison-Wesley.

Nancy Cartwright, 1983. *How the Laws of Physics Lie*. Oxford University Press.

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

R.G. Cowell, A.P. David, S.L. Lauritzen, and D.J. Spiegelhalter, 1999. *Probabilistic Networks and Expert Systems*. New York: Springer.

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 match-

ing 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.

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1998. Constructing partial prior specifications for models of complex physical systems. *The Statistician*, **47**, 37–53. With discussion.

C. Currin, T.J. Mitchell, M. Morris, and D. Ylvisaker, 1991. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

M. Goldstein, 1991. Belief transforms and the comparison of hypotheses. *Annals of Statistics*, **19**, 2067–2089.

M. Goldstein, 1999. Bayes linear analysis. In S. Kotz, editor, *Encyclopaedia of Statistical Sciences, update vol. 3*, pages 29–34. London: John Wiley & Sons.

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. forthcoming in the Journal of the American Statistical Association, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/BLCP.pdf`.

R. Haylock and A. O'Hagan, 1996. On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*, pages 629–637. Oxford, UK: Oxford University Press.

D. Higdon, M.C. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne, 2004. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, **26**(2), 448–466.

J.T. Houghton, Y. Ding, D.J. Griggs, M. Noguer, P. J. van de Linden, X. Dai, K. Maskell, and C.A. Johnson, editors, 2001. *Climate Change 2001: The Scientific Basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

M.C. Kennedy and A. O'Hagan, 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, **63**, 425–464. With discussion.

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.D. McKay, W. J. Conover, and R. J. Beckham, 1979. A comparison of three methods for selecting values of input variables in the analysis of output of computer code. *Technometrics*, **21**, 239–245.

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.

National Research Council (NRC), 1994. Report on statistics and physical oceanography. *Statistical Science*, **9**, 167–221. With discussion.

J.E. Oakley and A. O'Hagan, 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, **66**, 751–769.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J.E. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow, 2006. *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley.

A. O'Hagan, M.C. Kennedy, and J.E. Oakley, 1999. Uncertainty analysis and other inferential tools for complex computer codes. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 503–519. Oxford University Press. With discussion, pp. 520–524.

F. Pukelsheim, 1994. The three sigma rule. *The American Statistician*, **48**, 88–91.

C.S. Reese, A.G. Wilson, M. Hamada, H.F. Martz, and K.J. Ryan, 2004. Integrated analysis of computer and physical experiments. *Technometrics*, **46**(2), 153–164.

J.C. Rougier, 2006. Probabilistic inference for future climate using an ensemble of climate model evaluations. forthcoming in Climatic Change, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/CCfinal.pdf`.

J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, 1989. Design and analysis of computer experiments. *Statistical Science*, **4**(4), 409–423. With discussion, pp. 423–435.

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

M.L. Stein, 1989. Comment on "Design and analysis of computer experiments". *Statistical Science*, **4**(4), 432–433.

H. Stommel, 1961. Thermohaline convection with two stable regimes of flow. *Tellus*, **13**, 224–241.

G.W. Thomas, 1982. *Principles of Hydrocarbon Reservoir Simulation*. Boston: International Human Resources Development Corporation.

Bas C. van Fraassen, 1989. *Laws and Symmetry*. Clarendon Press.

K. Zickfeld, 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.

Table 5
Sensitivity analysis on the adjusted prediction for $F_1^{\text{crit}}$ from halving and doubling the various parameters (compare with the second row of Table 4).

| | Halving the value | | | Doubling the value | | |
|---|---|---|---|---|---|---|
| | Mean, $\mu$ | Sd dev., $\sigma$ | $(F_1-\mu)/\sigma$ | Mean, $\mu$ | Sd dev., $\sigma$ | $(F_1-\mu)/\sigma$ |
| | *Parameters affecting* $f \to f'$ | | | | | |
| $c_{ij}$ | 0.120 | 0.099 | −1.063 | 0.116 | 0.097 | −1.058 |
| $\mathsf{Sd}\,(\omega_{ij})$ | 0.120 | 0.096 | −1.108 | 0.114 | 0.109 | −0.919 |
| $\mathsf{Sd}\,(\nu_{ij})$ | 0.120 | 0.096 | −1.105 | 0.113 | 0.109 | −0.916 |
| $c_i$ | 0.118 | 0.099 | −1.057 | 0.119 | 0.099 | −1.062 |
| $\mathsf{Sd}\,(\delta_i(\cdot))$ | 0.118 | 0.098 | −1.061 | 0.119 | 0.100 | −1.050 |
| | *Parameters affecting* $f' \to f^*$ | | | | | |
| $\mathsf{Sd}\left(\omega_{ij}^*\right)$ | 0.128 | 0.082 | −1.388 | 0.103 | 0.143 | −0.624 |
| $\mathsf{Sd}\left(\nu_{ij}^*\right)$ | 0.120 | 0.096 | −1.110 | 0.114 | 0.109 | −0.915 |
| $c_i^*$ | 0.117 | 0.097 | −1.067 | 0.121 | 0.104 | −1.028 |
| $\mathsf{Sd}\,(\delta_i^*(\cdot))$ | 0.118 | 0.097 | −1.068 | 0.120 | 0.104 | −1.024 |
| | *Parameter affecting* $f^*(x^*, v^*) \to y$ | | | | | |
| $\mathsf{Sd}\,(\epsilon_i^*)$ | 0.120 | 0.091 | −1.172 | 0.115 | 0.126 | −0.804 |