Climate Simulators and Climate Projections

Jonathan Rougier[1] and Michael Goldstein[2]

[1]Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom; email: j.c.rougier@bristol.ac.uk

[2]Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, United Kingdom; email: michael.goldstein@durham.ac.uk

Annu. Rev. Stat. Appl. 2014. 1:103–23

First published online as a Review in Advance on November 20, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115652

Keywords

climate modeling, tuning, history matching, anomaly model, best-parameter model, model criticism

Abstract

We provide a statistical interpretation of current practice in climate modeling. In this review, we define weather and climate, clarify the relationship between simulator output and simulator climate, distinguish between a climate simulator and a statistical climate model, provide a statistical interpretation of the ubiquitous practice of anomaly correction along with a substantial generalization (the best-parameter approach), and interpret simulator/data comparisons as posterior predictive checking, including a simple adjustment to allow for double counting. We also discuss statistical approaches to simulator tuning, assessing parametric uncertainty, and responding to unrealistic outputs. We finish with a more general discussion of larger themes.

1. INTRODUCTION

Climate: your distribution of weather, represented as a multivariate spatiotemporal process (inherently subjective)

Weather: measurable aspects of the ambient atmosphere, notably temperature, precipitation, and wind speed

Our purpose in this review is to interpret current practice in climate modeling in the light of statistical inferences about past and future weather. In this way, we hope to emphasize the common ground between our two communities and to clarify climate modeling practices that may not, at first sight, seem particularly statistical. From this starting point, we can then suggest some relatively simple enhancements and identify some larger issues. Naturally, we have had to simplify many practices in climate modeling, but not—we hope—to the extent of making them unrecognizable.

1.1. Weather and Climate

We define weather to be measurable aspects of our ambient atmosphere, notably temperature, precipitation, and wind speed. Hence, weather is an objective property of the world. We define a climate to be a subjective distribution of weather, represented as a multivariate space-time stochastic process. The phrase "distribution of weather" is uncontentious, but we believe that much confusion has arisen from attempts to treat climate as an objective property of the world, rather than something associated with individuals that reflects their disposition to make bets—to adopt a simple operationalization of subjective probability, which we shall use throughout this review to treat all of the uncertainties within a common framework. Thus we write "your climate" rather than "the climate"; this specification seems to be the minimal change that is effective in emphasizing the subjective viewpoint.

This subjective definition of climate is not one that many climate scientists will recognize, and so we take a moment to evaluate it. First, one standard definition of climate is average weather, often represented as a 30-year arithmetic mean. Under this type of definition (which may be extended to a much richer summary), climate is an objective property of the world, being simply a known function of weather. Thus one could bet on climate, rather than, as we would have it, climate being the bet one makes on weather. We would then need a word for "distribution of climate" because climate has become synonymous with summaries of weather, so we have chosen to identify weather with its summaries and reserve climate for "distribution of weather."

What about the subjective element? Climate modelers may not be happy about statisticians telling them that the distribution of weather is subjective. They may, for example, point to the histogram of recent past weather as an objective distribution of weather, but a histogram is not a distribution. If you make the subjective judgment of temporal exchangeability, then the histogram of, e.g., 1980–2009 weather approximates your distribution of 2010 weather, albeit in a rather lumpy fashion. This approximation is because probabilistic updating of an exchangeable sequence implies convergence to the histogram. So climate modelers who share the judgment of exchangeability would roughly agree on the distribution of weather in 2010. But this argument is self-defeating because subjectivity is necessary to turn the histogram into a distribution. It also highlights a common mistake, which is to confuse agreement with nonsubjectivity.

The key point is that any probability for a unique event is unavoidably subjective (see, e.g., Hacking 2001). The weather event of at least one year of severe drought in England during 2020–2029 is unique: It cannot be embedded in a long sequence of similar but not identical events. At the moment, you may describe your assessment of this event probabilistically (it is an implication of your climate), but there is no reason to expect you to agree with anyone else. Your information, knowledge, and disposition are yours alone. Of course, a shared judgment of temporal exchangeability extending from 1980 to 2029 would be sufficient for agreement, but this

judgment is not defensible for a well-informed climate modeler, who is aware of the changes that are currently occurring in the earth system.[1]

Finally, our definition of climate seems to be consistent with current practice in climate modeling, as we describe in more detail in the following sections. A climate simulator is just that, a device for generating a family of distributions of weather. Insofar as the simulator is the outcome of many judgments, its distribution is subjective. Climate modelers do not accept one of the simulator's climates as their own but make a subjective adjustment reflecting their judgment about the simulator's limitations. So we find that the practice of climate modelers is inherently subjective, and that defining climate to be a subjective distribution of weather is reasonable not just from a foundational point of view but also from a naturalistic one.

Just to be absolutely clear, we use the word "subjective" solely to indicate that, by our definition, climate may vary from one person to another. When judgments are subjective, policymakers and the general public should exercise care when selecting their experts to ensure that these experts are qualified and representative. In the case of future weather, climate scientists are the experts, not statisticians like ourselves and not journalists or bloggers. A huge body of climate science is widely accepted within the climate science community, including that the net effect of human activity since 1750 has been one of warming (Solomon et al. 2007, Summary for Policymakers). These conclusions from the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report will shortly be reaffirmed in the Fifth Assessment Report (to be finalized in 2014).

Climate simulator: computer code that maps forcing into weather, given also an initial condition and parameter values

Forcing: boundary conditions for the simulator, usually implemented in terms of radiation (solar forcing and greenhouse gas forcing) and optical depth (aerosol forcing)

Statistical climate model: a statistical framework designed to simplify the process of specifying your climate

1.2. Climate Simulators and Their Uses

The earth can be represented as a forced system, driven by variations in isolation, by volcanism and other tectonic processes, and by human activity (Peixoto & Oort 1992). A climate simulator in its most primitive form is a function that maps forcings into weather, after which statistical postprocessing of the weather can be used to produce a climate; we discuss climate simulators in more detail in Section 2.1. We prefer the term "climate simulator" for the function that does the mapping, reserving "model" for its use (in the statistical sense) as a framework designed to simplify the process of specifying your climate (Rougier et al. 2013). But because the word "model" is heavily overloaded, we write either statistical climate model or XXX model, where XXX is the name we give to a particular statistical model.[2]

In our sense, statistical climate models will typically encompass climate simulators because the effect of forcing on the earth system has strong constraints that are induced by basic physical principles such as conservation and continuity. We can infer the qualitative effects of these constraints for a simplified earth; an example is the large-scale atmospheric organization known as Hadley cells (see, e.g., Ahrens 2000, chapter 11). However, a quantitative description on a realistic earth is less amenable to intuition and must be computed. Thus, quantitative statistical climate models are constructed in two stages: (*a*) develop a climate simulator that represents the physics and (*b*) propose a statistical model that represents your assessment of the simulator's limitations. Climate simulators are discussed in Section 2 and statistical climate models in Section 3.

[1] Exactly the same considerations apply to the weather of the past. Agreement about the climate of the recent past follows from the convergence of exchangeable judgments on the histogram. But where there is no histogram, for example for paleoweather, there is no particular reason for agreement.

[2] The term statistical climate model is also used for a class of statistical-dynamical simulators, in which a separation of scale argument is used to model the large-scale effects directly and to relegate the small-scale effects to a statistical ensemble (see Hasselmann 1976 for an outline of this approach and Petoukhov et al. 2000 for a description of the CLIMBER-2 statistical-dynamical simulator). We do not consider this type of simulator here.

Detection and attribution (D&A): examining the role of anthropogenic effects in 20th century weather patterns

Scenario: a future described by specified forcing, representative of trends in population, economics, technology, and policy interventions

Our distinction between "climate simulator" and "statistical climate model" is not widely made in climate science, but it exists implicitly because climate modelers use statistical models to adjust a simulator's climate. The ubiquitous model is that, for a quantity such as temperature, the simulator's climate is acceptable only up to an unknown fixed offset (the simulator bias), which in practice is estimated and plugged in. This is the statistical interpretation of anomaly correcting, in which, as a matter of course, a simulator's temperatures are vertically shifted so that the simulator mean temperatures during the period 1980–1999 exactly match the mean of observed temperatures during the same period (see the figures in Guttorp 2014). Statisticians will immediately see the opportunities for generalizing such a model. For example, the offset might be a spatiotemporal process; offsets from different types of output might be correlated; and rather than plugging in, the offset field might be integrated out.

We make this point right at the start of this review to stress that encompassing a climate simulator within a statistical climate model is not simply a statistician's conceit. Rather, it is something that already happens, but which, with statistical insight, could be generalized rather easily. Importantly, statistical judgments are necessary to move from climate simulator output to your climate, and these judgments must change as climate simulators evolve.

One role for climate simulators is hypothesis testing and predicting. Hypothesis testing is well illustrated by detection and attribution (D&A). In D&A, hypotheses compete to explain features such as the spatiotemporal structure of the warming trend in 20th century weather. Hypothesis A is that this trend is simply a realization of the weather's natural variability. Hypothesis B adds solar fluctuations and volcanism. Hypothesis C adds human activities. These hypotheses are statements about the forcing. To compute your likelihood ratio for, e.g., B versus C, you require your climate for a hypothetical earth without humans as well as your climate for the actual earth. An earth without humans can be implemented in a simulator by fixing the forcing from atmospheric greenhouse gases in the industrial period to be the same as that before the industrial period. Hypothesis tests for D&A are discussed in Rougier (2008a) and reviewed in Hegerl & Zwiers (2011); for reasons of space, they are not further covered here.

For predicting, policy interest is in future weather. In current practice, the future is represented in terms of scenarios for future forcing, which themselves arise from scenarios for population, economics, technology, and policy interventions.[3] Again, climate simulators provide the means of considering and comparing various hypothetical futures. They provide a platform for what-if intervention studies, such as geoengineering (e.g., Irvine et al. 2011), and for driving regional simulations for climate impact studies (Parry et al. 2007). Climate prediction is the main focus of this review, discussed in Sections 3 and 4.

2. CLIMATE SIMULATORS

In this review, we focus on large climate simulators, the state-of-the-art simulators that are run at the main climate research centers. The earth system comprises many interacting subsystems, most notably the atmosphere, hydrosphere, cryosphere, lithosphere, and biosphere, and the same is true of large climate simulators. Ahrens (2000) provides an introduction to weather (the companion volume, Stull 2000, is also helpful), whereas Peixoto & Oort (1992) provide a more mathematical treatment. McGuffie & Henderson-Sellers (2005) introduce climate modeling; Arakawa (1997) gives a technical treatment, outlining the mathematical issues involved in solving the underlying

[3] Williamson & Goldstein (2012) describe an adaptive approach to simulator-based policy assessment that avoids the use of scenarios.

equations; and Watanabe et al. (2010) describe some of the pragmatic choices that were made in constructing the MIROC5 simulator.

2.1. The Simulator as a Dynamical System

We do not consider the precise form of the laws governing the behavior and interactions of the simulator modules, beyond observing that the laws of the earth's subsystems are not all known and that they are not currently solvable at a scale sufficient to resolve all of the interesting processes. Instead, we focus on the nature of a climate simulator, which is a forced nonlinear dynamical system (see, e.g., McWilliams 2007, who also provides a useful overview of the challenges of climate modeling). Therefore, for a particular set of forcings (suppressed in the notation), we consider a climate simulator to be a deterministic function of time:

$$\mathbf{x}_t = \varphi(t; \mathbf{x}_0, \boldsymbol{\theta}), \quad \text{such that} \quad \varphi(t_0; \mathbf{x}_0, \boldsymbol{\theta}) = \mathbf{x}_0 \quad \text{for all } \boldsymbol{\theta},$$

where $\mathbf{x}_0$ is the initial climate state at time t_0.[4] The parameters $\boldsymbol{\theta}$ represent coefficients within the simulator code that are imperfectly known or that are too abstract to have an operational meaning; some parameters stand in for processes that are filtered out by the solver (sub–grid scale processes). Murphy et al. (2004) provide a list of approximately 30 such parameters, remarking that more than 100 are present in a typical large-scale simulator; we return to this idea in Sections 2.2 and 2.3.

The initial value, $\mathbf{x}_0$, must be supplied for the simulator to run, but that value presents a major problem in practice, being very high-dimensional and largely unknown even in the case in which the initial time, t_0, is contemporary. The difficulty is compounded if t_0 represents a historical initialization date such as 1850. The tendency in climate science has been to not specify $\mathbf{x}_0$ directly. Instead, the simulator, whose trajectories are chaotic, is treated as ergodic. A very long control run is made from a specified $\hat{\mathbf{x}}_0$ at a preferred set of parameter values (which we denote as $\tilde{\boldsymbol{\theta}}$ below) and with constant or periodic forcing; for example, the forcing of the year 1850 might be repeated again and again. For an ergodic simulator, time averages converge to the stationary measure, and hence an $\mathbf{x}_0$ for 1850, or a sequence of them, can be sampled from the control run after an interval of spin-up to forget $\hat{\mathbf{x}}_0$.

In this review, we do not consider $\mathbf{x}_0$ any further but focus instead on the role of $\boldsymbol{\theta}$, the simulator parameters. For simplicity, we focus on the expectation and variance of the simulator's climate, although other features (e.g., skewness and extremes) are also important for climate impact assessment. **Figure 1** summarizes the exposition of the next few paragraphs. These paragraphs represent a somewhat idealized interpretation of current practice, suggesting an opportunity for increased statistical sophistication, should climate modelers desire.

We write the simulator output at time t as $\mathbf{x}_t = \varphi(t; \boldsymbol{\theta})$ and the full set of simulator outputs as $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2, \ldots)$; $\mathbf{x}_t$ might itself represent a large collection of quantities, such as surface temperature, precipitation, and wind speed at every location on a 2° grid. The physics in the simulator suggests that, for each t, there will be strong relationships among the components of $\mathbf{x}_t$, and that these relationships will be somewhat consistent across t. The chaotic nature of the simulator suggests that $\mathbf{x}$ will look like a realization of a multivariate stochastic process, even when the forcing is smooth in time. For these two reasons, working directly with $\mathbf{x}$ is uncommon. Instead, a dimensionally reduced summary is used.

Parameters: adjustable coefficients in the simulator that represent sub–grid scale processes and incompletely understood processes

Sub–grid scale processes: processes in the mathematical model with length scales comparable to or smaller than the solver of the simulator

Control run: a long time-slice experiment, typically with preindustrial forcing

Spin-up: a time-slice experiment in which the simulator forgets its initial condition

[4]Here, we are simplifying by not distinguishing between the full state vector and the function of the state vector of interest to us, which would usually be a lower-dimensional summary. Once we dispense with the initial condition, treating $\mathbf{x}_t$ as the summary is sufficient.

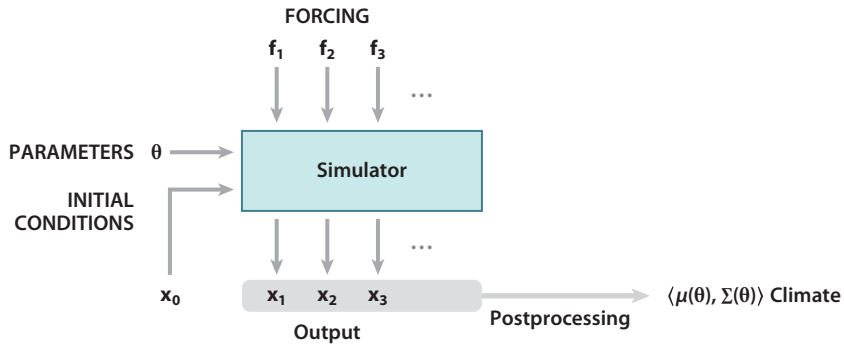

Figure 1

Schematic of a climate simulator. The inputs are forcing, $\mathbf{f} := (\mathbf{f}_1, \mathbf{f}_2, \ldots)$; parameter values $\boldsymbol{\theta}$; and an initial value $\mathbf{x}_0$. The simulator is run to produce outputs, $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2, \ldots)$. Using statistical time-series modeling, these inputs are summarized in terms of an expectation and a variance, which together represent the simulator's climate, collectively denoted $\mathcal{K}(\boldsymbol{\theta}) := \langle \mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}) \rangle$. The forcing should be included in the arguments of $\mathcal{K}$ but is suppressed for simplicity. The argument $\mathbf{x}_0$ is also suppressed, but in fact $\mathcal{K}$ should be nearly invariant to perturbations in $\mathbf{x}_0$, if the simulator is ergodic.

Let $\mathbf{x}$ be arranged as a matrix,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \end{bmatrix},$$

with singular value decomposition (SVD),

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{1}$ is the vector of ones and $\bar{\mathbf{x}}$ is the vector of column means (Golub & Van Loan 1996 describe the SVD and its properties in chapter 2). In climate modeling, the columns of $\mathbf{W} := \mathbf{U}\mathbf{D}$ are known as the empirical orthogonal functions (EOFs) (see, e.g., von Storch & Zwiers 1999, chapter 13); they are a bit like principal components, except that the rows of $\mathbf{X}$ are not exchangeable. For that reason, one should not confuse them. In this form, the simulator output can be written as

$$\mathbf{x}_t = \bar{\mathbf{x}} + \mathbf{V}\mathbf{w}_t \quad t = 1, 2, \ldots, \qquad 1.$$

where $\mathbf{w}_t^T$ is one row of $\mathbf{W}$. In practice, both $\mathbf{V}$ and $\mathbf{w}_t$ would be reduced from their full size to just the first k components, where k is determined empirically. In this case, it would be sensible to rescale the columns of $\mathbf{X}$ to be dimensionless before taking its decomposition, or else to apply dimensional reduction separately to each type of output.

Following dimensional reduction, the second step is to fit a time-series model to $\mathbf{w} := (\mathbf{w}_1, \mathbf{w}_2 \ldots)$. A simple choice would be to model the components of $\mathbf{w}_t$ independently, given that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and that $\mathbf{D}$ is diagonal, and to use a trend-plus-ARIMA (autoregressive integrated moving-average) residual model for each component (see, e.g., Chatfield 2004, chapter 4), including a seasonal component if the frequency is higher than annual. Including current and historical values of the forcing as covariates would also be legitimate and helpful, in which case the trend may be unnecessary. The time-series model for $\mathbf{w}$ can then be used to infer the second-order structure of $\mathbf{x}$ for a given $\boldsymbol{\theta}$ from Equation 1. We write the model here as

$$\mathbf{x} \sim \langle \mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}) \rangle,$$

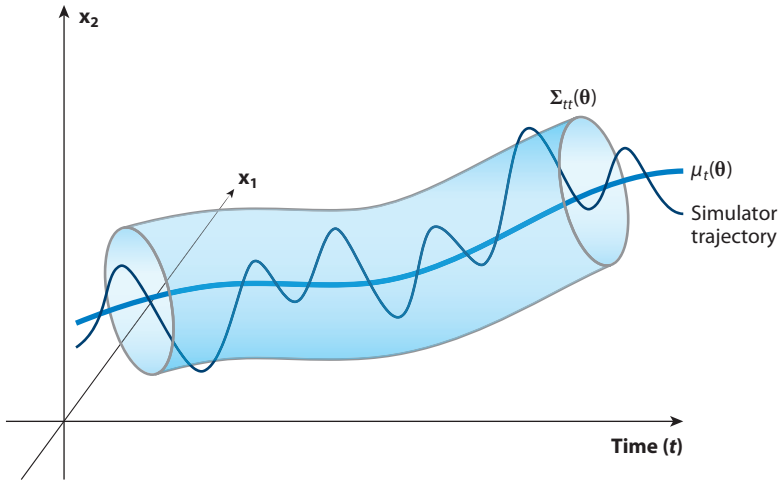

Figure 2

The set of simulator trajectories and the tube that summarizes the simulator's climate at parameter value θ.

where $\mu_t(\theta) := E(\mathbf{x}_t; \theta)$ and $\Sigma_{tt'}(\theta) := \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_{t'}; \theta)$. Technically, both μ and Σ should have hats to indicate that the parameters of the process for $\mathbf{w}$ have been estimated and then plugged in. Additional simulator runs with the same θ but different initial conditions can be used to improve the estimates of $\mu(\theta)$ and $\Sigma(\theta)$.

The tuple $\mathcal{K}(\theta) := \langle \mu(\theta), \Sigma(\theta) \rangle$ is synonymous with the climate of the simulator $\varphi(\cdot; \theta)$, at least to second order. $\mathcal{K}(\theta)$ describes an ellipse in the product space of time and the state, but for visualization purposes, we gain much by simplifying it to an elliptical tube in time, i.e., as a sequence $\{\mu_t(\theta), \Sigma_{tt}(\theta)\}$, which suppresses the temporal properties encoded in the covariances $\Sigma_{tt'}(\theta)$ (see **Figure 2**). The mean function, $\mu_t(\theta)$, can be used where a deterministic treatment of the simulator is required; typically it would be similar to a spatially and temporally smoothed version of the raw output. $\Sigma_{tt}(\theta)$ is termed the internal variability of the simulator output at time t. If the time-series model for the residual of $\mathbf{w}$ is stationary, then $\Sigma_{tt}(\theta)$ is invariant to t, and the tube has the same shape all the way along.

In the special case of a time-slice experiment, the forcing is constant or periodic, and the simulator is run until its output stabilizes. Typically, the mean of the final 30 years of the run is used, suppressing the internal variability to the point at which it can be ignored. Time-slice experiments are used to summarize the climate of different epochs (e.g., the preindustrial era, the Last Glacial Maximum) and to compute hypothetical quantities such as equilibrium climate sensitivity. They are also used for tuning (see Sections 2.2 and 2.3).

For reasons of computational scale, it is sensible to perform inferences in the feature space, $\mathbf{W} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T)\mathbf{V}$, rather than in the original space, but in this article we stay with the original space for simplicity. Rather than take $\bar{\mathbf{x}}$ and $\mathbf{V}$ from the actual run, one may instead take them from the control run; in this case, $\mathbf{w}$ might be modeled with a multivariate time-series model.

Internal variability: the inherent variability of the weather within a climate simulator

Time-slice experiment: experiment in which the simulator is given constant or periodic forcing and run at least until its output stabilizes

Tuning: choosing a preferred value for the simulator's parameters

2.2. Tuning

Tuning is the activity of choosing a preferred value for θ, which we denote as $\tilde{\theta}$. The major constraint of tuning is the slow integration time of the simulator. Large-scale simulators can typically compute approximately 100 simulator years per calendar month. To tune to a long time

series, such as a century of regional temperature and precipitation, would additionally require a spin-up for each new choice of parameter values, involving at least another 100 simulator years. And so an iterative process with this target would move at approximately one cycle for every two calendar months, which is hopeless when there are hundreds of uncertain parameters. Therefore climate simulators are, in general, not extensively tuned to reproduce the large-scale features of 20th century weather, especially as tuning generally happens in the shadow of a looming IPCC deadline. As a result, 20th century weather can be used to assess climate model adequacy (see Section 4).

Overtuning: tuning that emphasizes a match with the histogram of 20th century weather

This observation should be tempered, though, in the light of the sequential development of simulators within a research group. Many of the decisions made when upgrading a climate simulator are based on increased computer power or better physical understanding, but some are based on the failure of the current version of the simulator to reproduce 20th century weather.[5] Valdes (2011) notes that climate simulators are currently too stable to replicate historical abrupt weather transitions. This problem might be symptomatic of overtuning, with the Holocene (our current epoch) being unusually stable. One solution is also to tune on previous epochs with different forcing. The difficulty here is that the forcings are much more uncertain and the histogram of weather must be inferred from proxy measurements (see, e.g., Jones et al. 2009).

Two camps exist regarding tuning strategies: (*a*) that the modules of the climate simulator should be tuned separately, so as to avoid compensatory mistuning and (*b*) that climate is an emergent property of the interactions of its subprocesses, and so tuning should happen jointly. Typically, the camps reach something of a compromise. Gent et al. (2011, section 3) summarize the procedure for CCSM4 (Community Climate System Model).[6] First, the modules of the simulator (atmosphere, ocean, land, sea ice) were each separately tuned to reproduce current behavior. Danabasoglu et al. (2008) illustrate the combination of physical and empirical reasoning that is used to tune one aspect of the ocean module.[7] Module tuning uses both time-slice experiments (to check for long-run stability) and also transient runs, during which observations replace the fluxes from the other modules.

Module tuning takes care of most of the parameters. Then Gent et al. (2011) coupled the modules together into a climate simulator, and the simulator as a whole was tuned on a few parameters and a few targets. A cloud parameter was adjusted to achieve a satisfactory radiation balance at the top of the atmosphere (TOA), and sea-ice albedo parameters were adjusted to give satisfactory sea-ice thicknesses in the Arctic.

Gent et al. (2011) also summarize the diagnostic evaluation of CCSM4 at the tuned value $\tilde{\theta}$, using observed 20th century weather. Crucially, this evaluation is not just in terms of mean fields, e.g., for temperature and precipitation, although those get checked first, but also in terms of the statistical properties of variability (e.g., the histogram of precipitation) and recurrent events [e.g., the El Niño–Southern Oscillation (ENSO)]. In other words, the purpose of tuning the simulator is not simply to get $\mu(\tilde{\theta})$ approximately right but also to get key features of $\Sigma(\tilde{\theta})$ approximately right.

In our experience, the procedure for CCSM4 is unusually ascetic, with most modeling groups tuning jointly on a larger set of parameters and a larger set of targets. Mauritsen (2012) provides a detailed description of the process of tuning the MPI-ESM (Max Planck Institute–Earth System

[5]The genealogy of climate simulators is highly instructive (see Knutti et al. 2013 and Masson & Knutti 2011).

[6]This is an open-source climate simulator, which makes it rather unusual, as climate simulators are typically proprietary to climate modeling groups. CCSM4 has now been subsumed under CESM1 (see **http://www.cesm.ucar.edu/models/ccsm4.0/**).

[7]Tuning involves much more than just adjusting parameters; often the parameter set itself is changed as chunks of code are swapped.

Model) simulator. Public descriptions of the practice of tuning a large climate simulator are a recent phenomenon.

2.3. History Matching

History matching (HM): statistical approach for ruling out poor choices of the simulator parameters

Projection: a climate prediction along a specified scenario

Emulator: statistical model for a simulator, allowing prediction of the simulator output at untried values of the parameters

Statisticians have lots of tools to help with the process of tuning a climate simulator. Here we outline an exploratory approach termed history matching (HM), which demands much less expert judgment than does fully probabilistic conditioning.[8] HM is designed to rule out bad choices for the parameter values. Not-ruled-out values—for which the simulator outputs are consistent with historical observations—do not necessarily have similar simulator outputs under different forcings (such as in future projections). The intention of HM, in contrast to tuning, is to preserve this source of climate uncertainty. HM originated in hydrocarbon reservoir modeling and is used extensively in commercial settings. Its original statistical formulation appeared in Craig et al. (1997). Vernon et al. (2010) provide a detailed description of HM for a galaxy simulator, Gladstone et al. (2012) for the Pine Island Glacier, Edwards et al. (2011) (termed precalibration) for an intermediate-complexity climate simulator, and McNeall et al. (2013) for an ice-sheet simulator.

As an illustration, we take just a single target for tuning, the TOA mean radiation balance in an 1850 time-slice experiment, with any value θ with an imbalance outside the range $(-0.1, 0.1)$ W/m^2 being deemed unacceptable as a candidate for the preferred value (see, e.g., Gent et al. 2011, p. 4,977). The width of this target interval should include a component for tolerability (large discrepancies being tolerable for some targets but not for others) and also for measurement error (see, e.g., Vernon et al. 2010, section 3.5). HM inverts this constraint to rule out regions of the parameter space. To proceed efficiently, it exploits the property that the simulator output for the target is a smooth deterministic function of the parameters. So in this case, the simulator output would be, e.g., a 30-year mean of the radiation imbalance from the end of the run, denoted $\bar{\mu}(\theta)$, and internal variability can be neglected.

The simulator is treated as an unknown smooth deterministic function of the parameters (or a subset of them), represented by a statistical model termed an emulator.[9] An emulator is a sophisticated response surface, typically containing both regressors for global effects and a stochastic process for local effects. A catalog of carefully chosen runs at different points in the parameter space is used to update the emulator, and the result is an expectation and a standard deviation for $\bar{\mu}(\theta)$ at any θ. Note that $\bar{\mu}(\theta)$ is a random quantity even though θ is specified, if the simulator has not been run at θ. In the special case in which the simulator has been run at θ, to give, e.g., a value v, then the smoothness of the emulator ensures that $\mathrm{E}\{\bar{\mu}(\theta)\} = v$ and $\mathrm{SD}\{\bar{\mu}(\theta)\} = 0$. Various strategies exist for choosing the set of runs, but a popular initial choice is a Latin hypercube. Santner et al. (2003) and Forrester et al. (2008) provide more details about emulation and experimental design. Rougier (2008b) develops an emulator for multivariate outputs, such as a spatial field (see Rougier et al. 2009a for an illustration).

Based on the emulator, any point in the parameter space can be scored according to whether the predicted value overlaps with the target. Thus, a particular choice θ might be deemed unacceptable as a candidate for the preferred value if the intersection between $(-0.1, 0.1)$ and $\mathrm{E}\{\bar{\mu}(\theta)\} \pm 3 \times \mathrm{SD}\{\bar{\mu}(\theta)\}$ is not empty.

Any θ for which the intersection is empty is not ruled out yet (NROY). Vernon et al. (2010) give a detailed description of the process of HM with an emulator and how it can proceed in successive

[8] Sansó et al. (2008) and Tokmakian & Challenor (2013) provide examples of fully probabilistic calibration.

[9] Emulators are required only for expensive simulators. Gladstone et al. (2012), for example, use the simulator directly in the HM procedure.

waves, allowing more and more of the parameter space to be ruled out through additional runs of the simulator and refittings of the emulator. Vernon et al. (2010) also discuss HM with multiple targets and low-dimensional visualizations of an implausibility measure defined on the parameter space.

Anomaly model: a statistical climate model in which the dominant limitation of the simulator is mislocation of your climate

Best-parameter model: a generalization of the anomaly model, in which the limitations of the simulator involve the location, size, and shape of your climate

2.3.1. Fast approximate simulators. The emulation approach is particularly powerful for simulators for which fast approximations exist. For climate simulators, these simulators would typically have lower resolution, with prognostic variables replaced by diagnostic variables (i.e., removing feedback from some of the state variables) or with shorter spin-ups. For this approach to be effective, running the simulator in fast mode must be relatively easy, and this mode must be designed in from the start.

For example, emulators can include arbitrary smooth functions of the parameters as regressors, and the fast approximate simulators (FASs) could be one such. In this way, one can think of an emulator as a statistical approach to correcting an FAS. After the emulator is built, which requires paired runs of both the full simulator and the FAS, the parameter space can then be explored at the speed of the FAS. Intuitively, if the FAS is a poor approximation that is difficult to correct, then not much of the parameter space will be ruled out because $\mathrm{SD}\{\bar{\mu}(\theta)\}$ will tend to be large. In practice, a more sophisticated use of emulators is possible, linking simulators through the coefficients in their emulators (see Cumming & Goldstein 2009). In climate science, Rougier et al. (2009b) use a FAS to provide prior information for the HadSM3 climate simulator, and Williamson et al. (2012) link the low-resolution FAMOUS (FAst Met Office UK Universities Simulator) climate simulator with the high-resolution HadCM3 (Hadley Centre Coupled Model, version 3) simulator.

3. STATISTICAL CLIMATE MODELS

In this section, we make an important transition from the climate simulator, thought of as a function of the parameters θ, to your climate. In moving from one to the other, we pass into the realm of subjective judgments, as explained in Section 1.1. If this subjectivity is not obvious in current practice, it must be concealed in the widespread acceptance of conventional judgments. And, indeed, the anomaly model described in Section 3.1 is exactly this: a conventional judgment for passing from a simulator's climate to your climate, thus concealing the essential subjectivity of this step.

Conventions can be supremely useful, of course. One example is the symmetry-breaking convention of agreeing to drive on the left-hand side of the road (in the United Kingdom). Conventional simplifications, however, must continually be reappraised as our understanding and our tools develop. Thus, many of the conventional simplifications in climate modeling have now been relaxed, and parts of the earth system previously ignored or treated as diagnostic have become prognostic (e.g., the sulfur cycle, vegetation). The anomaly model is a conventional simplification of judgments that goes back to the very start of climate modeling, and the time has come to relax it, too. For example, the so-called best-parameter model that we discuss in Section 3.2 is a useful first step in this direction.

In this section, we contrast two different approaches to inferences about future weather under particular future forcings, termed climate projections. The time domain is divided into the past ($t \in \mathcal{P}$), for which observations exist, and the future ($t \in \mathcal{F}$), for which we would like to make a projection. For the time being, we treat the past forcing as known (but see Section 4.2); the future forcing is specified by the projection scenario. For concreteness, $\mathcal{P}$ might be the period 1850–2013 and $\mathcal{F}$ the period 2014–2100, and the simulator output might be annual global mean

temperature. The time period $\mathcal{P}$ does not have to be contiguous with that of $\mathcal{F}$, and for simplicity, we take $\Sigma_{\mathcal{P}\mathcal{F}}(\theta) = \mathbf{0}$ for each θ.[10]

Let $\mathbf{Y} = [\mathbf{Y}_{\mathcal{P}}, \mathbf{Y}_{\mathcal{F}}]$ be the weather, and let

$$\mathbf{z}^{\text{obs}} = \mathbf{Y}_{\mathcal{P}} \oplus \mathbf{e}$$

be the statistical model for past weather observations, where $\mathbf{e}$ is the measurement error with expectation zero and known variance matrix $\mathbf{E}$, and $\oplus$ indicates the addition of uncorrelated components.[11] Then the objective is to make inferences about $\mathbf{Y}_{\mathcal{F}}$ based on $\mathbf{z}^{\text{obs}}$ and on runs of the simulator. We restrict ourselves to the single simulator run at the preferred parameter value, $\tilde{\theta}$, represented in terms of the simulator's climate $\mathcal{K}(\tilde{\theta}) = \langle \mu(\tilde{\theta}), \Sigma(\tilde{\theta}) \rangle$. Section 3.3 discusses the important issue of alternative choices for θ.

3.1. The Anomaly Model

Our statistical interpretation of climate modelers' current behavior is that they take a classically frequentist approach to climate inference, proposing a strongly parametric statistical model linking the climate simulator and their climate, estimating the parameters of this statistical model, and then plugging in the estimated values to make probabilistic projections. We term the climate modelers' current statistical model the anomaly model. It asserts the existence of parameters (θ^*, α^*) with the property that

$$\mathbf{Y} | \theta^*, \alpha^* \sim \left\langle \mu(\theta^*) + \alpha^* \mathbf{1}, \Sigma(\theta^*) \right\rangle, \qquad 2.$$

where α^* is a scalar anomaly correction. This statistical model asserts that $\mathcal{K}(\theta^*)$ adequately represents your climate, except for an unknown translation. Equation 2 is the simplest version and can be easily generalized to allow the anomaly correction to have multiple components, which depend on the type of weather output and possibly the spatial location. Some outputs may need to be transformed so that an additive shift is appropriate (e.g., precipitation, which is nonnegative on its natural scale).

The tuned value $\tilde{\theta}$ is taken to be an estimate of θ^*, and α^* is then estimated as

$$\tilde{\alpha} = (n_p)^{-1} \sum_{t \in \mathcal{P}} \left(\mathbf{z}_t^{\text{obs}} - \mu_t(\tilde{\theta}) \right) \qquad 3.$$

in the simplest version, where n_p is the number of time points in the period $\mathcal{P}$. Thus the anomaly correction is a function of $\mathbf{z}^{\text{obs}}$, even if $\tilde{\theta}$ is not. One then finds the projection by plugging in the estimate $(\tilde{\theta}, \tilde{\alpha})$ for the unknown (θ^*, α^*) to give

$$\mathbf{Y}_{\mathcal{F}} \sim \left\langle \mu_{\mathcal{F}}(\tilde{\theta}) + \tilde{\alpha} \mathbf{1}, \Sigma_{\mathcal{F}\mathcal{F}}(\tilde{\theta}) \right\rangle. \qquad 4.$$

The preceding discussion is a rather long-winded way of saying, "Having found, by other means, a preferred value for θ, translate the simulator climate so that it matches, on average, the

[10]In practice, a much more detailed set of outputs would be used—generalizing the approach described below is straightforward. Where $\mathcal{P}$ and $\mathcal{F}$ are contiguous, $\Sigma_{\mathcal{P}\mathcal{F}}(\theta) \approx \mathbf{0}$ is implied by short (e.g., not more than a decade) correlation lengths in the residual component of the time-series model for $\mathbf{w}$. Again, generalizing is straightforward.

[11]We are skipping over the nature of these observations. Many weather observations start out as indirect measurements, e.g., from weather satellites, which measure radiances at different wavelengths that are then processed ("inverted") to give temperatures. Common sources of uncertainty in the forward relationship from temperature to radiance would induce systematic errors in the observations (see Section 4.2), but it would be unusual for the weather and the measurement error to be correlated.

historical observations."[12] But we emphasize that the parametric model in Equation 2 is a subjective assessment of the relationship between the simulator's climate and your climate, notwithstanding the aura of objectivity that arises from the apparent absence of any explicit quantification of uncertainty. We return to this idea in Section 3.2.

Discrepancy: in the best-parameter model, the uncertain additive difference between the simulator output at its best parameterization and your climate

3.2. The Best-Parameter Model

The anomaly model of Section 3.1 is rather simple. The statistical field of computer experiments has developed a richer set of models for simulator-based inference for complex systems (see, e.g., Craig et al. 2001; Goldstein & Rougier 2004, 2006; Kennedy & O'Hagan 2001); Rougier (2007) provides a climate science illustration. The approach in computer experiments is based around a so-called best-parameter statistical model, which in this context asserts that there exists a $\boldsymbol{\theta}^*$ such that $\mathcal{K}(\boldsymbol{\theta}^*)$ is second-order sufficient for your climate. This model can be written as

$$\mathbf{Y}|\boldsymbol{\theta}^*, \boldsymbol{\varepsilon} \sim \langle \mu(\boldsymbol{\theta}^*) + \boldsymbol{\varepsilon},\ \Sigma(\boldsymbol{\theta}^*) \rangle,$$

where $\boldsymbol{\varepsilon}$ is an additive discrepancy, probabilistically independent of $\boldsymbol{\theta}^*$, with $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{m}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{T}$ and with both $\mathbf{m}$ and $\mathbf{T}$ being specified. We write "discrepancy" instead of "anomaly correction" because $\boldsymbol{\varepsilon}$ is a random vector, not a scalar shift. Taking $\mathbf{m} = \mathbf{0}$ is common because known translations can be incorporated directly into the simulator, and we follow this practice from now on. Integrating out $\boldsymbol{\varepsilon}$ gives

$$\mathbf{Y}|\boldsymbol{\theta}^* \sim \langle \mu(\boldsymbol{\theta}^*), \Sigma(\boldsymbol{\theta}^*) + \mathbf{T} \rangle.$$

The discrepancy variance $\mathbf{T}$ would capture your remaining uncertainty about climate, were you able to run the simulator at its best parameter. First, you would not expect $\mathcal{K}(\boldsymbol{\theta}^*)$ to be in quite the right place, and so there should be a term that allows for translations, like the anomaly correction. Second, $\mathcal{K}(\boldsymbol{\theta}^*)$ is typically underdispersive with respect to your climate, owing to the limited resolution of the solver, which acts as a filter.[13] Thus, in the simplest version of the best-parameter model,

$$\boldsymbol{\varepsilon} = \alpha^* \mathbf{1} \oplus \boldsymbol{\varepsilon}_2, \quad \text{implying} \quad \mathbf{T} = \sigma_1^2 \mathbf{1}\mathbf{1}^T + \mathbf{T}_2, \tag{5.}$$

where $\alpha^* \sim \langle 0, \sigma_1^2 \rangle$, and $\mathbf{T}_2 := \mathrm{Var}(\boldsymbol{\varepsilon}_2)$ might be as simple as the diagonal matrix, $\sigma_2^2 \mathbf{I}$. But $\mathbf{T}$ also has the capacity to encode changes in the shape of $\mathcal{K}(\boldsymbol{\theta}^*)$ to give a richer and more appropriate description of the simulator's discrepancies at different time points.

In this model, the projection is made by updating using the historical observations, giving

$$\mathbf{Y}_{\mathcal{F}}|\boldsymbol{\theta}^*, z^{\mathrm{obs}} \sim \langle \boldsymbol{\mu}_{\mathcal{F}|\mathcal{P}}(\boldsymbol{\theta}^*), \boldsymbol{\Sigma}_{\mathcal{F}|\mathcal{P}}(\boldsymbol{\theta}^*) \rangle,$$

where, taking $\boldsymbol{\Sigma}_{\mathcal{P}\mathcal{F}} = \mathbf{0}$ and suppressing $\boldsymbol{\theta}^*$ (which is an argument for all terms of the form $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$),

$$\begin{aligned} \boldsymbol{\mu}_{\mathcal{F}|\mathcal{P}} &:= \boldsymbol{\mu}_{\mathcal{F}} + \mathbf{T}_{\mathcal{F}\mathcal{P}}(\boldsymbol{\Sigma}_{\mathcal{P}\mathcal{P}} + \mathbf{T}_{\mathcal{P}\mathcal{P}} + \mathbf{E})^{\dagger}(\mathbf{z}^{\mathrm{obs}} - \boldsymbol{\mu}_{\mathcal{P}}) \\ \boldsymbol{\Sigma}_{\mathcal{F}|\mathcal{P}} &:= \boldsymbol{\Sigma}_{\mathcal{F}\mathcal{F}} + \mathbf{T}_{\mathcal{F}\mathcal{F}} - \mathbf{T}_{\mathcal{F}\mathcal{P}}(\boldsymbol{\Sigma}_{\mathcal{P}\mathcal{P}} + \mathbf{T}_{\mathcal{P}\mathcal{P}} + \mathbf{E})^{\dagger}\mathbf{T}_{\mathcal{P}\mathcal{F}}. \end{aligned}$$

These are the usual second-order updating equations (Goldstein & Wooff 2007, Rougier et al. 2013), where † denotes the Moore–Penrose inverse.[14] For a plug-in projection, $\boldsymbol{\theta}^*$ can be replaced by its estimate, $\tilde{\boldsymbol{\theta}}$.

[12] In fact, as mentioned in Section 1, the convention is to match the means over the period 1980–1999 rather than the whole of $\mathcal{P}$, as $\mathcal{P}$ might differ from one experiment to another.

[13] This is not the only source of underdispersion: Simplifications are also present in the modeling of processes such as sea ice and vegetation, reflecting both computational constraints and lack of knowledge.

[14] They are also the conditioning expressions for the multivariate Gaussian distribution.

Interestingly, this best-parameter model includes the anomaly model as a special case. We illustrate in the simplest version, given in Equation 5. If

$$\sigma_1^2 \mathbf{1}\mathbf{1}^T \gg \mathbf{T}_2 \quad \text{and} \quad \sigma_1^2 \mathbf{1}\mathbf{1}^T \gg \Sigma_{\mathcal{P}\mathcal{P}} + \mathbf{E}, \qquad 6.$$

Overfitting: a consequence of using the anomaly model, resulting in too-small projection uncertainties

then $(\Sigma_{\mathcal{P}\mathcal{P}} + \mathbf{T}_{\mathcal{P}\mathcal{P}} + \mathbf{E})^\dagger \approx \mathbf{T}^\dagger_{\mathcal{P}\mathcal{P}} = n_p^{-2}\sigma_1^{-2}\mathbf{1}\mathbf{1}^T$. Simple arithmetic then shows that the projection is the same as in the anomaly model: $\mu_{\mathcal{F}|\mathcal{P}} = \mu_{\mathcal{F}} + \tilde{\alpha}\mathbf{1}$ and $\Sigma_{\mathcal{F}|\mathcal{P}} = \Sigma_{\mathcal{F}\mathcal{F}}$, where $\tilde{\alpha}$ was defined in Equation 3. The assertions in Equation 6 state that your concern for the mislocation of the simulator's climate dominates all other uncertainties. A climate modeler who does not believe Equation 6 would regard the anomaly-corrected projection Equation 4 as overfitted. This modeler would attribute part of the systematic difference between $\mathbf{z}^{\text{obs}}$ and $\mu_{\mathcal{P}}(\theta^*)$ to internal variability and so adjust the location of the simulator climate by less and decrease the projection uncertainty by less.

Compared with the anomaly model, the best-parameter model allows much more flexibility for the discrepancy, including that it is a stochastic process. For example, $\alpha_t^* = \rho\alpha_{t-1}^* + \eta_t$, with $\alpha_0^* \sim \langle 0, \sigma_1^2 \rangle$ and η a sequence of uncorrelated innovations with $\eta_t \sim \langle 0, (1-\rho^2)\sigma_1^2 \rangle$, for which the anomaly model is the special case of $\rho = 1$. This stochastic process for α_t^* allows the anomaly to be both uncertain and time varying, with the value ρ controlling the temporal correlation length. The main difficulty for climate modelers is that it seems more acceptable to specify $\rho = 1$ rather than choose a value for ρ that is less than one. Likewise, specifying $\sigma_2 = 0$ after Equation 5 seems more acceptable than choosing a value greater than zero.[15] Although many more detailed judgments can be represented in $\mathbf{T}$, if the starting point is the anomaly model, then $\rho < 1$ and $\sigma_2 > 0$ are very simple extensions.

Sexton et al. (2012) is the one study in climate modeling that has explicitly included a full discrepancy variance. The authors use an approach based on an ensemble of simulators from different modeling groups, generating one realization of the discrepancy per simulator and estimating $\mathbf{T}$ from the result. The few simulators in their ensemble would produce a severely rank-deficient value for $\mathbf{T}$, but Sexton et al. perform the entire inference in the much lower-dimensional feature space (see Section 2.1). Although we are hesitant to endorse this particular approach, our view is that any approach that helps climate modelers propose a $\mathbf{T}$ that adjusts both the location and the dispersion of the simulator's climate is welcome, provided that the results are not inconsistent with the modelers' judgments.

3.3. Uncertainty About the Best Parameter

The projections in Sections 3.1 and 3.2 were made for a specific parameter value, the preferred value $\tilde{\theta}$, thought of as a plug-in point estimate for θ^*. That $\tilde{\theta}$ is not θ^*, and that θ^* remains uncertain, is a concern in climate modeling, and several experiments have assessed the sensitivity of simulator output to parameter perturbations (reviewed by Murphy et al. 2011). So far, though, we have seen no attempt to quantify the effect of parameter uncertainty on climate projections for the current generation of climate simulators.

Formally, this quantification is straightforward: $\mathcal{K}(\theta)$ denotes the expectation and variance conditional on $\theta^* = \theta$, and so integrating out θ^* gives

$$\mathcal{K}^* = \langle \mathrm{E}[\mu(\theta^*)], \mathrm{E}[\Sigma(\theta^*)] + \mathrm{Var}[\mu(\theta^*)] \rangle. \qquad 7.$$

[15] Boundary values such as $\rho = 1$ and $\sigma_2 = 0$ give the appearance of objectivity but are of course just as subjective as any other values and less defensible than many. Box (1980, p. 384) commented on "the curious idea that an outright assumption does not count as a prior belief."

Finite approximations can replace the expectations and variance. In the simplest case, these approximations would be from an ensemble of runs sampled from the prior distribution for θ^* (Rougier 2007), but more sophisticated approaches are possible using emulators (Craig et al. 2001, Rougier & Sexton 2007, Rougier et al. 2009b). The expectation in $\mathcal{K}^*$ will not be equal to $\mu(\tilde{\theta})$, even in the case where $\tilde{\theta}$ is the prior expectation of θ^*, because μ is a nonlinear function, perhaps extremely so in some parts of the parameter space (McWilliams 2007). And $\Sigma(\tilde{\theta})$ will typically understate the variance in $\mathcal{K}^*$ because of its missing the $\mathrm{Var}[\mu(\theta^*)]$ term.

However, exploring the effect of uncertainty in θ^* comes with a computational price because today's climate simulators are approximately as large as computing resources will allow, and there are no spare CPU cycles for replication. So running different candidate values for θ^* is possible only by reducing the simulator's resolution.[16] Halving the resolution of today's simulators would allow approximately ten low-resolution simulator runs instead of one high-resolution simulator run, putting aside concerns about spinning up.

In their choices, climate modelers currently reveal a strong preference for one high-resolution run, rather than, e.g., ten low-resolution runs.[17] This reluctance to use low-resolution runs for assessing projection uncertainty has implications for the statistical model. The replacement of θ^* by the plug-in point estimate $\tilde{\theta}$ could be compensated by increasing the variance of the discrepancy in the best-parameter model, i.e., letting $\mathbf{T}_2$ in Equation 5 represent $\mathrm{Var}[\mu(\theta^*)]$. But this compensation would require an explicit departure from the anomaly model, for which $T_2 \approx \mathbf{0}$, and so climate modelers find themselves at a statistical impasse. As we perceive it, the choice to do one high-resolution run instead of a set of low-resolution runs is incompatible with the use of the anomaly model to link the climate simulator and actual climate, unless one is prepared to defend the judgment that $\mathrm{Var}[\mu(\theta^*)] \approx \mathbf{0}$, i.e., that perturbing the parameters has a negligible effect on the expectation of the simulator's climate.

We also believe that the reluctance to perform low-resolution runs is misplaced. Crudely, today's low-resolution simulator φ_{low} is the previous IPCC report's state-of-the-art simulator. At the time of the previous report, perturbations of φ_{low} were thought to be informative about your climate. Today, perturbations of φ_{high} are thought to be informative about your climate. Accepting the premise—which is not disputed—that climate simulators are currently much more like one another than any one climate simulator is like your climate, you must accept that perturbations in φ_{low} are informative about perturbations in φ_{high}. Goldstein & Rougier (2009) discuss statistical models for sequences of simulators, and Rougier et al. (2013) discuss statistical models for a collection of climate simulators of roughly equal fidelity.

We hope that by the time of the sixth IPCC report, climate modelers will be exploring the limitations of the tuning process and quantifying the effect of parametric uncertainty. Ideally, this exploration would take the form of carefully designed experiments combining runs of low- and high-resolution simulators, which we believe represents the most efficient way to exploit a fixed budget of CPU cycles (Cumming & Goldstein 2009).

[16] Or by other approaches to speeding up the simulator, as discussed in Section 2.3, but here we focus on resolution.

[17] We will not speculate on why this is so. But as (statistical) modelers ourselves, we are familiar with the vexed issue of realism. After a successful upgrade, today's climate simulator looks appreciably more realistic than before. For example, Gent et al. (2011, section 5d) document the improvement of CCSM4 over CCSM3 in representing ENSO. ENSO is an important feature of the earth system and a milestone for climate simulation, and it must be painful for climate modelers who have finally achieved this milestone to then reduce resolution and lose it again. Salt (2008) provides an interesting reflection on modeling culture.

4. MODEL CRITICISM

4.1. Computing Residuals

Model criticism, also termed "model validation" by engineers, attempts to evaluate whether the statistical climate model is adequate for its purpose. There is a lot of informal model criticism in current climate modeling, but, at first glance, very little formal statistical model criticism. Informal model criticism tends to proceed by comparing $\mathbf{z}^{\text{obs}}$ with the anomaly-corrected simulator output $\mu_{\mathcal{P}}(\tilde{\theta}) + \tilde{\alpha}\mathbf{1}$ and seeing whether the differences are large relative to the standard deviations in $\Sigma_{\mathcal{P}\mathcal{P}}(\tilde{\theta}) + \mathbf{E}$ (see the statement in Randall et al. 2007, section 8.1.2.3). Interestingly, under the anomaly model, this approach is precisely the posterior predictive checking (PPC) approach originally proposed by Rubin (1984) and advocated by Gelman et al. (2003, chapter 6).

In the PPC approach, one notionally replicates the observations, in our case giving rise to the second-order Directed Acyclic Graph

$$\begin{array}{ccccc} \phi(\mathcal{P} : \theta^*) & \longrightarrow & \mathbf{y}_{\mathcal{P}} & \longrightarrow & \mathbf{z}^{\text{obs}} \\ & & \searrow & & \\ & & & & \mathbf{z}^{\text{rep}}, \end{array}$$

where $\mathbf{z}^{\text{rep}}$ are the notionally replicated observations. Then $\mathbf{z}^{\text{obs}}$ is evaluated with respect to the expectation and variance of $\mathbf{z}^{\text{rep}}|\mathbf{z}^{\text{obs}}$. Under the anomaly model, now thought of as a special case of the best-parameter model as discussed in Section 3.2,

$$\mathbf{y}_{\mathcal{P}}|\theta^*, \mathbf{z}^{\text{obs}} \sim \left\langle \mu_{\mathcal{P}}(\theta^*) + \tilde{\alpha}\mathbf{1}, \Sigma_{\mathcal{P}\mathcal{P}}(\theta^*) \right\rangle,$$

and the result then follows by taking $\mathbf{z}^{\text{rep}} = \mathbf{y}_{\mathcal{P}} \oplus \mathbf{e}^{\text{rep}}$, where $\mathbf{e}^{\text{rep}}$ is uncorrelated with $\mathbf{e}$ but has the same expectation and variance, and by plugging in $\tilde{\theta}$ for θ^*.

Gelman et al. (2003) advocate using graphical summaries of $\mathbf{z}^{\text{obs}}$ in the distribution of $\mathbf{z}^{\text{rep}}|\mathbf{z}^{\text{obs}}$; in climate, these might be plots of rescaled prediction errors,

$$r_t := \frac{z_t^{\text{obs}} - (\mu_t(\tilde{\theta}) + \tilde{\alpha})}{\sqrt{\Sigma_{tt}(\tilde{\theta}) + E_{tt}}} \quad t \in \mathcal{P}, \qquad 8.$$

which can also be very effective if $\mathbf{z}^{\text{obs}}$ is a spatial map. These residuals are not standardized to have variance one when the model is adequate because of the double counting of $\mathbf{z}^{\text{obs}}$, which is used to estimate $\tilde{\alpha}$. The adjustment is straightforward. If $\mathbf{H} := \mathbf{I} - (n_p)^{-1}\mathbf{1}\mathbf{1}^T$ (the centering matrix; Mardia et al. 1979, chapter 1), then

$$\mathbf{z}^{\text{obs}} - (\mu_{\mathcal{P}}(\tilde{\theta}) + \tilde{\alpha}\mathbf{1}) = \mathbf{H}(\mathbf{z}^{\text{obs}} - \mu_{\mathcal{P}}(\tilde{\theta})),$$

and so, letting $\mathbf{V} := \Sigma_{\mathcal{P}\mathcal{P}}(\tilde{\theta}) + \mathbf{E}$, the denominator of the residuals in Equation 8 should be not $\sqrt{V_{tt}}$ but $\sqrt{(HVH)_{tt}}$. Slight modifications to $\mathbf{H}$ would be required for different anomaly conventions (see Footnote 12). Multivariate information including covariances can be harder to visualize. Bastos & O'Hagan (2009) give a simple approach based on the pivoted Cholesky decomposition of the predictive variance.

4.2. Diagnostic Warnings

How should the climate modeler respond to a subset of diagnostics that are large in absolute size?[18] Suppose, for example, that many of the standardized precipitation residuals are larger than three

[18]We can also imagine situations in which some linear combinations of the residuals are surprisingly small because the climate model has failed to respect physical constraints. However, at the moment, climate modelers are mainly concerned with residuals that are too large.

in absolute size in a region such as western Europe, where precipitation changes are an important feature of climate change impact. Several options exist:

1. Acknowledge that the simulator does not yet "do" precipitation and request a more powerful computer, hoping that higher resolution will reduce the residuals;
2. Acknowledge that the simulator has been badly tuned and restart the tuning process, hoping that a better choice for $\tilde{\theta}$ will reduce the residuals;
3. Acknowledge that the anomaly model is a rather simplistic representation of judgments about the simulator and your climate, hoping that a better specification for the discrepancy variance T will reduce the residuals.[19]

In all three options, climate modelers should decline to make projections for precipitation, given the failure of the model. Instead, they should wait for a new computer, or for the simulator to be retuned, or while they refine their judgments about $\mathbf{T}$.

We would strongly recommend exploring option 3 first. Once $\mathbf{T}$ is explicitly specified, the climate modeler can compute prior predictive residuals

$$r'_t := \frac{z_t^{\text{obs}} - \mu_t(\tilde{\theta})}{\sqrt{\Sigma_{tt}(\tilde{\theta}) + T_{tt} + E_{tt}}} \quad t \in \mathcal{P} \qquad 9.$$

(Box 1980). These are standardized to have expectation zero and variance one when the model is adequate.

The idea that $\mathbf{T}$ might be adjusted retrospectively to improve the residuals $\mathbf{r}'$ needs to be clearly motivated. First, at a pragmatic level, the climate modeler might have decided to use the anomaly model if possible, despite it not being a defensible representation of the modeler's judgments. Thus, large residuals under the anomaly model indicate that $\mathbf{T}$ is an area of the inference in which the modeler must make an additional effort.

Second, the modeler's adjustment should be to the whole of $\mathbf{T}$, not just to the submatrix $\mathbf{T}_{\mathcal{P}\mathcal{P}}$. The modeler might, for example, reflect on whether $\rho = 1$ and $\sigma_2 = 0$ were really appropriate choices (see Section 3.2). Setting $\rho < 1$ and/or $\sigma_2 > 0$ will increase both the historical uncertainty about $\mathbf{z}$ and the projection uncertainty about $\mathbf{Y}_{\mathcal{F}}$. It is self-evident that a failure of the simulator to match historical observations increases uncertainty about climate projections. Adjusting the whole of $\mathbf{T}$ to improve the residuals in $\mathcal{P}$ is a simple implementation that has the effect of increasing prediction uncertainty to reflect a poor fit to observations.

Third, from its position in Equation 9, $\mathbf{T}$ can clearly take on some of the burden of specifying $\Sigma_{\mathcal{P}\mathcal{P}}(\tilde{\theta})$ and $\mathbf{E}$. Both of these terms are challenging. $\Sigma_{\mathcal{P}\mathcal{P}}(\tilde{\theta})$ is expressed for a specified forcing, yet 20th century forcing is not well known (Forster et al. 2007). So some of the increase in $\mathbf{T}$ might reflect the modeler's assessment of uncertainty in the forcing. For $\mathbf{E}$, Guttorp (2014) describes some of the issues with climate observations. Statisticians will appreciate that common uncontrolled sources of variation in the collection and processing of instrumental readings introduce nonzero off-diagonal elements into the observation error variance, $\mathbf{E}$; no attempt has been made, so far, to tackle this challenge. So, again, some of the covariances in $\mathbf{T}$ might reflect common sources of variation in the observations.

Finally, conceptual difficulties exist in the best-parameter model, particularly in specifying judgments that are coherent for a sequence of simulators. Goldstein & Rougier (2009) introduce a more general approach, termed reified modeling. Thus, although modifying $\mathbf{T}$ might be the

[19] A more sophisticated variant of this option is to use statistical downscaling to adjust the simulator climate (see Maraun et al. 2010).

incremental response to poor residuals, the reified modeling approach might turn out to be a better representation of a climate modeler's actual judgments.

5. SUMMARY AND PROSPECTS

Two issues should engage statisticians working in climate research. First, given where we are, what is the pragmatic statistical analysis that will best complement/enhance current climate practice? Second, given an ideal position, what is the analysis that we would like to carry out to best inform us about future weather?

This review is largely concerned with aspects of the first issue. We briefly summarize the main sources of uncertainty that we have identified for a given climate projection:

1. Input uncertainty, which is not knowing the historical and future boundary conditions;
2. Parametric and structural uncertainty, which arise from limitations in the climate simulator (and subsume other uncertainties such as the effect of code errors and numerical noise);
3. Observational error, including common components that induce covariances in the observation error variance;
4. Code uncertainty, which is being unable to run the simulator at every desired parameter setting.

These sources of uncertainty are all familiar to statisticians but not routinely addressed in climate projections. We hope that our review has identified some simple extensions of current practice and also the opportunity for more detailed treatment, where judgments allow.

We now turn to the second issue. Changes in weather over the present century have the capacity to threaten literally hundreds of millions of people. To note just one peril, Nicholls (2011) discusses the effect of sea-level rise: Currently, more than 200 million people are vulnerable to flooding during extreme storms, and the probability of a catastrophe will increase as sea levels continue to rise through the century. Among the huge uncertainties affecting the risk of flooding are the behavior of the Greenland and West Antarctic ice sheets, the intensification of tropical and extratropical storms, and changes to surge propagation (ibid., pp. 147–148). Our society's response to this peril and others like it might come to be seen as one of the defining features of political and social activity for the current century.

Addressing the original four uncertainties is clearly already a large challenge for climate modelers. But when we consider the impact of the weather, we have to allow for additional uncertainties, principally:

5. Downscaling uncertainty, which maps from the large grid cell of global climate simulators to the small grid cell that is necessary to evaluate losses, taking account of local topography and bathymetry;
6. Loss uncertainty, which is valuing the harm and damage caused by climate-related hazards;
7. Decision uncertainty, which is the uncertain consequence of an intervention, which in turn depends on social and economic factors.

All of these items are single-simulator concerns, so we must add the following:

8. Multisimulator uncertainty, accounting for the sequence of simulators within each research group and the different simulators across research groups.

Now a full uncertainty assessment for climate policy appears doubly daunting. However, that is the wrong way to look at the problem. The really daunting task is to make and successfully implement a climate policy without doing a careful uncertainty analysis.

Although these sources of uncertainty are challenging to assess, they are no more challenging than other parts of climate modeling (e.g., doing the basic science, formulating mathematical

models, constructing simulators that run efficiently on supercomputers, and designing satellite missions to collect observations). The difference is that the climate modeling challenges are addressed by well-established communities. If climate policy is a genuine concern, then scientifically leading countries such as the United Kingdom need to develop a similar community of climate statisticians, working alongside the other communities and funded at a sufficient level, with the same access to computing facilities.

In this case, we would hope to see rapid development along the lines outlined in this review. This development would include the replacement of tuning with history matching, the incorporation of parametric uncertainty into projections, and the use of FASs in experimental design and emulation. In a few years, we would hope to see a gradual acceptance among climate modelers that the distribution of weather is subjective, and that current approaches based on suppressing that subjectivity using boundary choices for statistical parameters, as in the anomaly model, are indefensible. We will need creative approaches for specifying the discrepancy variance matrix. Accordingly, we hope for the development of more powerful statistical modeling approaches for linking multiple simulators to put climate policy at the heart of the inference (our own suggestion is reified modeling; see Goldstein & Rougier 2009).

Within the decade, we would look for the development of new statistical techniques that modularize inference for future weather impacts in the same way that climate modeling itself is modularized. We would expect these techniques to be based on dynamical graphical models, for which the main challenge is the very high level of interconnectivity among the vertices. These developments in statistical methodology would be widely applicable, and useful for any complex systems in which uncertainty about real-world consequences is informed by families of computer simulators.

Finally, we address the political dimension of the issues that we have discussed. If we knew precisely what the future had in store, then considering what perils we should protect against, and to what degree, would be relatively easy. However, our world is far too complex to offer such certainty. A common line of argument is that the element of subjectivity involved in climate projections justifies taking a wait-and-see attitude toward actions for climate change mitigation and adaptation. This argument has a potentially paralyzing effect on informed discussion over climate policy, and scientists, understandably concerned that such arguments will be used to discount the utility of their assessment, may be tempted to downplay the uncertainty associated with their projections.

This temptation is doubly unfortunate. First, downplaying or objectifying uncertainty makes much valuable and informative work in climate science an easy target for groups with a vested interest in preserving the status quo (see, e.g., Oreskes & Conway 2010). Second, this suppression of uncertainty masks much of the case for taking action now because climate projections are by no means worst-case scenarios. Rather, they are located near the center of a range of possible future climates, all of which should be considered to develop appropriate climate policies. Such policies will, inevitably, be constructed in a context of uncertainty, and this uncertainty can be assessed only in terms of expert judgments, based on a synthesis of all the available evidence.

That some disagreement exists among scientists studying climate is natural and inevitable. However, as mentioned in the Introduction, the broad lines of the argument for human-induced climate change are clear (and we should not forget that this peril is just one of many we face; see Rockström et al. 2009). Rational choice of action in complex problems requires careful consideration of both uncertainties and consequences. Analysis of the most careful and detailed climate projections that are possible within current computational constraints consistently suggests the potential for human activity to lead to disastrous real-world consequences. A plurality of expert judgments about probabilities and consequences does not diminish the case for acting now.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank James Annan, Philip Brohan, Richard Chandler, Peter Challenor, Mat Collins, Tamsin Edwards, Lindsay Lee, Reto Knutti, and Danny Williamson for very helpful comments on an earlier draft of this review and absolve them completely of any responsibility for the views expressed herein. The margin definitions given in this article are by necessity very crude. In some cases, more details are given in the text; in most, we advise consulting a standard source such as the IPCC glossary or the World Meteorological Organization glossary.

LITERATURE CITED

Ahrens CD. 2000. *Meteorology Today: An Introduction to Weather, the Climate, and the Environment*. Pacific Grove, CA: Brooks/Cole. 6th ed.

Arakawa A. 1997. Adjustment mechanisms in atmospheric models. *J. Meteorol. Soc. Jpn.* 75(1B):155–79

Bastos LS, O'Hagan A. 2009. Diagnostics for Gaussian process emulators. *Technometrics* 51(4):425–38

Box GEP. 1980. Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. A* 143(4):383–430

Chatfield C. 2004. *The Analysis of Time Series*. Boca Raton, FL: Chapman & Hall/CRC

Craig PS, Goldstein M, Rougier JC, Seheult AH. 2001. Bayesian forecasting for complex systems using computer simulators. *J. Am. Stat. Assoc.* 96:717–29

Craig PS, Goldstein M, Seheult AH, Smith JA. 1997. Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case Studies in Bayesian Statistics*, Vol. III, ed. C Gatsonis, JS Hodges, RE Kass, RE McCulloch, P Rossi, ND Singpurwalla, pp. 37–87. New York: Springer-Verlag

Cumming JA, Goldstein M. 2009. Small sample Bayesian designs for complex high-dimensional models based on information gained using fast approximations. *Technometrics* 51(4):377–88

Danabasoglu G, Ferrari R, McWilliams JC. 2008. Sensitivity of an ocean general circulation model to a parameterization of near-surface eddy fluxes. *J. Clim.* 21:1192–208

Edwards N, Cameron D, Rougier JC. 2011. Precalibrating an intermediate complexity climate model. *Clim. Dyn.* 37:1469–82

Forrester AIJ, Sóbester A, Keane AJ. 2008. *Engineering Design via Surrogate Modelling: A Practical Guide*. Chichester, UK: John Wiley & Sons

Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, et al. 2007. Changes in atmospheric constituents and in radiative forcing. See Solomon et al. 2007, pp. 129–234

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.

Gent PR, Danabasoglu G, Donner LJ, Holland MM, Hunke EC, et al. 2011. The Community Climate System Model version 4. *J. Clim.* 24:4973–91

Gladstone RM, Lee V, Rougier JC, Payne AJ, Hellmer H, et al. 2012. Calibrated prediction of Pine Island Glacier retreat during the 21st and 22nd centuries with a coupled flowline model. *Earth Planet. Sci. Lett.* 333–34:191–99

Goldstein M, Rougier JC. 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM J. Sci. Comput.* 26(2):467–87

Goldstein M, Rougier JC. 2006. Bayes linear calibrated prediction for complex systems. *J. Am. Stat. Assoc.* 101:1132–43

Goldstein M, Rougier JC. 2009. Reified Bayesian modelling and inference for physical systems (with discussion). *J. Stat. Plan. Inf.* 139:1221–56

Goldstein M, Wooff DA. 2007. *Bayes Linear Statistics: Theory & Methods*. Chichester, UK: John Wiley & Sons
Golub GH, Van Loan CF. 1996. *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press. 3rd ed.
Guttorp P. 2014. Statistics and climate. *Annu. Rev. Stat. Appl.* 1:87–101
Hacking I. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge, UK: Cambridge Univ. Press
Hasselmann K. 1976. Stochastic climate models. Part I: theory. *Tellus* 28:473–85
Hegerl GC, Zwiers FW. 2011. Use of models in detection and attribution of climate change. *WIREs Clim. Change* 2:570–91
Irvine PJ, Ridgwell A, Lunt DJ. 2011. Climatic effects of surface albedo geoengineering. *J. Geophys. Res.* 116:D24112
Jones PD, Briffa KR, Osborn TJ, Lough JM, van Ommen TD, et al. 2009. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *Holocene* 19(1):3–49
Kennedy MC, O'Hagan A. 2001. Bayesian calibration of computer models (with discussion). *J. R. Stat. Soc. B* 63:425–64
Knutti R, Masson D, Gettelman A. 2013. Climate model genealogy: generation CMIP5 and how we got there. *Geophys. Res. Lett.* 40:1194–99
Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, et al. 2010. Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* 48:RG3003
Mardia KV, Kent JT, Bibby JM. 1979. *Multivariate Analysis*. London: Academic
Masson D, Knutti R. 2011. Climate model genealogy. *Geophys. Res. Lett.* 38:L08703
Mauritsen T, Stevens B, Roeckner E, Cruger T, Esch M, et al. 2012. Tuning the climate of a global model. *J. Adv. Model. Earth Syst.* 4:M00A01
McGuffie K, Henderson-Sellers A. 2005. *A Climate Modelling Primer*. Chichester, UK: John Wiley & Sons. 3rd ed.
McNeall DJ, Challenor PG, Gattiker JR, Stone EJ. 2013. The potential of an observational data set for calibration of a computationally expensive computer model. *Geosci. Model Dev. Discuss.* 6:2369–401
McWilliams JC. 2007. Irreducible imprecision in atmospheric and oceanic simulations. *Proc. Natl. Acad. Sci. USA* 104(21):8709–13
Murphy JM, Clark R, Collins M, Jackson C, Rodwell M, et al. 2011. Perturbed parameter ensembles as a tool for sampling model uncertainties and making climate projections. *Proc. ECMWF Workshop Represent. Model Uncertain. Error Numer. Weather Clim. Predict. Models, 20–24 June 2011*, pp. 183–208. Reading, UK: Eur. Cent. Medium-Range Weather Forecast. **http://www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/2011/Model_uncertainty/Murphy.pdf**
Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, et al. 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 430:768–72
Nicholls RJ. 2011. Planning for the impacts of sea-level rise. *Oceanography* 24(2):144–57
Oreskes N, Conway EM. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury
Parry ML, Canziani OF, Palutikof JP, van der Linden PJ, Hanson CE, eds. 2007. *Climate Change 2007: Impacts, Adaptation and Vulnerability*. Cambridge, UK: Cambridge Univ. Press
Peixoto JP, Oort AH. 1992. *Physics of Climate*. New York: Springer
Petoukhov V, Ganapolski A, Brovkin V, Claussen M, Eliseev A, et al. 2000. CLIMBER-2: a climate system model of intermediate complexity. Part I: model description and performance for present climate. *Clim. Dyn.* 16:1–17
Randall DA, Wood RA, Bony S, Colman R, Fichefet T, et al. 2007. Climate models and their evaluation. See Solomon et al. 2007, pp. 589–662
Rockström J, Steffen W, Noone K, Persson Å, Chapin FS III, et al. 2009. A safe operating space for humanity. *Nature* 461:472–75
Rougier JC. 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim. Change* 81:247–64
Rougier JC. 2008a. Annotated bibliography: climate change detection and attribution. *ISBA Bull.* 15(4):3–6. **http://bayesian.org/sites/default/files/fm/bulletins/0812.pdf**

Rougier JC. 2008b. Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Stat.* 17(4):827–43

Rougier JC, Goldstein M, House L. 2013. Second-order exchangeability analysis for multi-model ensembles. *J. Am. Stat. Assoc.* 108:852–63

Rougier JC, Guillas S, Maute A, Richmond A. 2009a. Expert knowledge and multivariate emulation: the thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics* 51(4):414–24

Rougier JC, Sexton DMH. 2007. Inference in ensemble experiments. *Philos. Trans. R. Soc. A* 365:2133–43

Rougier JC, Sexton DMH, Murphy JM, Stainforth D. 2009b. Analysing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *J. Clim.* 22(13):3540–57

Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12(4):1151–72

Salt JD. 2008. The seven habits of highly defective simulation projects. *J. Simul.* 2:155–61

Sansó B, Forest C, Zantedeschi D. 2008. Inferring climate system properties using a computer model (with discussion). *Bayesian Anal.* 3(1):1–62

Santner TJ, Williams BJ, Notz WI. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer

Sexton DMH, Murphy J, Collins M, Webb MJ. 2012. Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Clim. Dyn.* 38(11–12):2513–42

Solomon S, Qin D, Manning M, Chen Z, Marquis M, et al., eds. 2007. *Climate Change 2007: The Physical Science Basis.* Cambridge, UK: Cambridge Univ. Press

Stull RB. 2000. *Meteorology for Scientists and Engineers*. Pacific Grove, CA: Brooks/Cole. 2nd ed.

Tokmakian R, Challenor P. 2013. Uncertainty in modeled upper ocean heat content change. *Clim. Dyn.* In press. doi:10.1007/s00382-013-1709-9

Valdes P. 2011. Built for stability. *Nat. Geosci.* 4:414–16

Vernon I, Goldstein M, Bower RG. 2010. Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Anal.* 5(4):619–70

von Storch H, Zwiers FW. 1999. *Statistical Analysis in Climate Research*. Cambridge, UK: Cambridge Univ. Press

Watanabe M, Suzuki T, O'ishi R, Komuro Y, Watanabe S, et al. 2010. Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. *J. Clim.* 23:6312–35

Williamson D, Goldstein M. 2012. Bayesian policy support for adaptive strategies using computer models for complex physical systems. *J. Oper. Res. Soc.* 63:1021–33

Williamson D, Goldstein M, Blaker A. 2012. Fast linked analyses for scenario-based hierarchies. *Appl. Stat.* 61(5):665–91