ANNUAL REVIEWS

Annual Review of Statistics and Its Application

On the Statistical Formalism of Uncertainty Quantification

James O. Berger[1] and Leonard A. Smith[2,3]

[1]Department of Statistical Science, Duke University, Durham, North Carolina 27708, USA; email: berger@duke.edu

[2]Centre for the Analysis of Time Series, Department of Statistics, London School of Economics, London WC2A 2AE, United Kingdom

[3]Pembroke College, University of Oxford, Oxford OX1 1DW, United Kingdom; email: lenny@maths.ox.ac.uk

Annu. Rev. Stat. Appl. 2019. 6:433–60

First published as a Review in Advance on December 3, 2018

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

https://doi.org/10.1146/annurev-statistics-030718-105232

Keywords

uncertainty, predictability, prediction, forecast, insight, simulation, weather-like, climate-like, discrepancy, uncertainty quantification, uncertainty guidance, emulation

Abstract

The use of models to try to better understand reality is ubiquitous. Models have proven useful in testing our current understanding of reality; for instance, climate models of the 1980s were built for science discovery, to achieve a better understanding of the general dynamics of climate systems. Scientific insights often take the form of general qualitative predictions (i.e., "under these conditions, the Earth's poles will warm more than the rest of the planet"); such use of models differs from making quantitative forecasts of specific events (i.e. "high winds at noon tomorrow at London's Heathrow Airport"). It is sometimes hoped that, after sufficient model development, any model can be used to make quantitative forecasts for any target system. Even if that were the case, there would always be some uncertainty in the prediction. Uncertainty quantification aims to provide a framework within which that uncertainty can be discussed and, ideally, quantified, in a manner relevant to practitioners using the forecast system. A statistical formalism has developed that claims to be able to accurately assess the uncertainty in prediction. This article is a discussion of if and when this formalism can do so. The article arose from an ongoing discussion between the authors concerning this issue, the second author generally being considerably more skeptical concerning the utility of the formalism in providing quantitative decision-relevant information.

1. INTRODUCTION

Modeling nature is a long-standing goal in science. The availability of computers has drastically increased the amount of such modeling. Quantifying the uncertainty in model-based statements is a fundamental aspect of science; in weather and in climate, this has long been recognized explicitly. After World War II, weather prediction was selected as the prime nonclassified test case for large-scale modeling with electronic computers. In 1979, the Charney Report (NRC 1979) called for the use of computer models to better inform decisions regarding anthropogenic climate change. In the weather context, the phrase "No forecast is complete without an estimate of its uncertainty" has become a rallying cry for those implementing operational ensemble prediction systems (EPS) (Tennekes et al. 1986). Since 1992, operational weather forecasting has included EPS in both Europe and the United States; while this Monte Carlo approach provides an estimate of the sensitivity of a point forecast given the state of the atmosphere today (Molteni et al. 1996), it remains unclear just how the distributions obtained are best interpreted. Another modeling application is to adjudicate between competing theories or the manner in which an event comes to be; for example, models can distinguish between two different theoretical ideas to establish how (model) tornadoes develop by providing insight into the dynamics that we are not currently able to observe directly (Parker 2014).

Weather modeling is clearly a predictive success story. On October 13, 2013, the plausibility of a major storm on October 28 (St. Jude's Day) was indicated by weather models, 15 days in advance. The prior probability of a storm of this magnitude on this date near this location would be very small. Six days in advance, the storm appeared in many of the 51 operational simulations, and three days in advance, virtually every simulation showed a major storm. In this sense, weather models can be attributed longer-range foresight than any human forecaster. We refer to these tasks as "weather-like," noting that uncertainty quantification (UQ) for weather-like tasks has access to a huge forecast-outcome data archive.

In contrast, situations in which the lead time of the forecast is long compared with the lifetime of a model (or even a modeler) are called "climate-like." There are also attempts to use computer models to predict details of a distant reality, say, the temperature of the hottest day in 2094; neither this goal nor quantifying the uncertainty of today's attempts to do so is plausible, for reasons that will be discussed. Providing clear guidance on whether the uncertainty quantified is regarding something in the real world, or merely for some aspect of the next model run, is a crucial part of UQ.

The interface of the scientific modeling enterprise with data and uncertainty is called uncertainty quantification in the math/engineering world. In statistics, this name appears redundant, as the primary role of statistics is uncertainty quantification; in economics, the phrase is an oxymoron, as Knightian uncertainty is by definition unquantifiable. Nevertheless, UQ has become the standard name for the interface, and we follow that custom.

Many elements of UQ are highly statistical, so involvement of statistics in the area is crucial. The first goal of this article is to outline the main statistical issues involved with UQ, highlighting those areas in which statisticians can play an important role. Central to this goal is the clarification of precisely what UQ aims to deliver.

The second goal of the article is to discuss a statistical formalism that has developed, following in the steps of Kennedy & O'Hagan (2001), that claims to be able, in principle, to accurately assess the uncertainty in predictions from models. The current article arose from a long ongoing discussion of the two authors concerning the extent to which this claim holds true, with the first author generally supportive of the formalism and the second quite skeptical of its ability to provide decision-relevant probabilities for geophysical systems; the article lands somewhere between

these two positions. Wherever the truth falls, we agree that statisticians can play a larger role in maintaining the credibility of large-scale computer modeling and that this is a task of importance in both decision making and policy support (Smith & Stern 2011, Parker & Risbey 2015).

Examples that we use to illustrate the key issues include geophysical models such as climate and weather models (on timescales of a day, week, season, and century), models of volcanic pyroclastic flows, and models of the behavior of car suspension systems (the last being a small enough problem that we can actually describe the entire statistical UQ process). A statistical discussion of UQ in geophysical models can be found in Smith (2014) and references therein. There is also, of course, an extensive parallel effort to do such modeling using purely statistical models of processes, and, indeed, we will see that some such modeling is necessary for effective use of computer models.

2. THE STATISTICAL FORMALISM FOR UQ

The interface of statisticians and UQ is explored in Sacks et al. (1989), Currin et al. (1991), Welch et al. (1992), and Morris et al. (1993). General discussions of the entire validation and verification process for computer models can be found in Roache (1998), Oberkampf & Trucano (2000), Easterling (2001), Pilch et al. (2001), and Trucano et al. (2002). Starting with Kennedy & O'Hagan (2001) (see also Craig et al. 1997), the formalism described in this section has been developed over many publications; overviews include Kennedy & O'Hagan (2000), Craig et al. (2001), Goldstein & Rougier (2003, 2004), Santner et al. (2003), Bayarri et al. (2007a, 2007b, 2009b), Gramacy & Lee (2009), Ba & Joseph (2012), Berliner (2003), and Smith (2014).

2.1. Notation for the Formalism

The key elements of the formalism as seen from a statistical perspective are as follows. Many of these are standard statistical modeling issues, but at least one is not.

1. **Real process:** $y^R(\mathbf{x})$ is the output of the real process of interest, the system. The evolution of the system in time varies with the input $\mathbf{x}$. For instance, $y^R(\mathbf{x})$ could be the maximum height of volcanic pyroclastic flow at a location, depending on $\mathbf{x} =$ (volume, direction) of the initial flow. In other contexts, inputs $\mathbf{x}$ could be state variables that evolve in time or physical parameters (e.g., the speed of light, the melting point of pure lead) that are thought to be constant in time. It can be important to distinguish these two varieties of input.
2. **Computer model output:** $y^M(\mathbf{x}, \mathbf{u})$ is the computer simulation model created to predict the real process. For the pyroclastic flow problem, this would typically be a computer implementation of a partial differential equation (PDE) model based on finite elements and the physics of flow. In addition, computer models typically have calibration parameters, $\mathbf{u}$, which play a critical role in the model yet may or may not have any physical or mathematical meaning in the context of the real process.
3. **Model discrepancy (bias):** $b(\mathbf{x}, \mathbf{u}) = y^R(\mathbf{x}) - y^M(\mathbf{x}, \mathbf{u})$. Computer models arguably never provide exact predictions of the real process (excepting computations restricted to small integers). The difference between reality and the computer model prediction is called the discrepancy (or bias), when it is well-defined. Discrepancy results from a variety of sources, including incorrect and missing physics (e.g., the interaction of a pyroclastic flow with the ground can be only crudely modeled), limited realism of the boundary conditions (e.g., the topography of the island is not modeled to the level of detail at which it is observed), and effects of casting the problem into a digital framework.

The presence of discrepancy is a departure from common statistical practice, wherein the fitting of a statistical model to data is thought to mostly eliminate structural model error and thus bias. Of course, it is recognized that statistical models are also typically wrong in some way, but statistical practice typically proceeds by trying to identify when a model is inadequate, and then utilizing a better model. This effort to improve a model is also a significant part of UQ; nevertheless, one is often required to work with the formal presence of a discrepancy term.

4. **Observations of the real process:** $y^O(\mathbf{x}) = y^R(\mathbf{x}) + \epsilon$, where ϵ is any kind of additive statistical measurement error. This is the ordinary formulation for a statistical model of observations from the real process; it is not universally applicable.
5. **Prior probability distributions:** The quantities $\mathbf{x}, \mathbf{u}, b(\cdot,\cdot), \epsilon$ are all typically at least partly unknown and are assigned probability distributions. It is not clear what the prior on a meaningless $\mathbf{u}$ represents.
6. **(Possibly) an emulator of $y^M(\mathbf{x}, \mathbf{u})$:** Often, each simulation by the computer model is quite expensive (2 hours for the pyroclastic flow computer model); state-of-the-art weather forecasting models often use all the available computing time before the next forecast is required. It can be advantageous to create an approximation to the computer model. While this may prove operationally challenging, it is conceptually straightforward for statistics:
   - View the computer model, $y^M(\mathbf{x}, \mathbf{u})$, as a function to be modeled and approximated.
   - Run the computer model over some designed set of inputs $D = \{(\mathbf{x}_i, \mathbf{u}_i), i = 1, \ldots, n\}$.
   - Define the emulator as the predictive distribution for $y^M(\cdot)$ resulting from a statistical model $y^E(\mathbf{x}, \mathbf{u}) = y^M(\mathbf{x}, \mathbf{u}) + \eta(\mathbf{x}, \mathbf{u})$, where $\eta(\mathbf{x}, \mathbf{u})$ is some statistical random error; the most commonly used error process is a Gaussian process, for which spatial statistics methodology can be used to construct the emulator.
7. **Analysis:** Perform Bayesian analysis with the above entities, obtaining the predictive posterior distribution, $p(y^R(\mathbf{x}))$, of reality at any input $\mathbf{x}$ (and the posterior distributions of any other quantities of interest, such as the calibration parameters). The advantage of doing this in a Bayesian fashion is that the accuracy assessments of all predictions and parameter estimates are available (if the computation can be handled), accounting for all of the (modeled) uncertainties.

This statistical formalism of UQ is based on some assumptions that may prove problematic in applications, and, worse, there is often no way of knowing that all uncertainties have been addressed or that a given prediction task can actually be placed in the framework. We discuss these problems in Section 4.

2.2. Case Study: Vehicle Suspension System

Bayarri et al. (2007a) did a UQ analysis for a computer model of vehicle suspension systems, and we review the main steps in the analysis here. (Not every step of the UQ analysis is given, since some are too technical for the presentation here.)

2.2.1. The elements of the uncertainty quantification analysis.

- **Real process:** The real process considered was that of driving a vehicle over a test road with two major potholes and recording the resulting forces on the suspension system.
  —$\mathbf{x} = (x_1, \ldots, x_7)$ is the input vector of key suspension system characteristics.

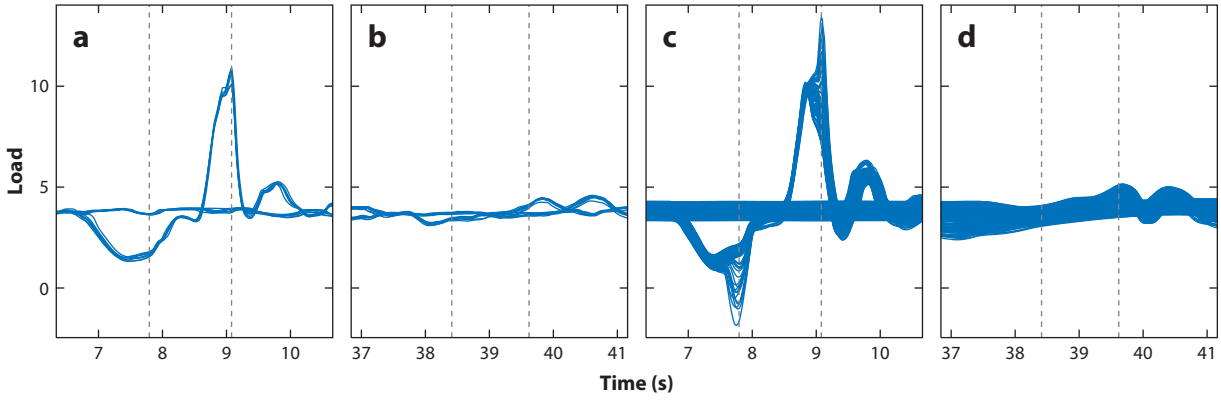

Figure 1

Forces on one location on the suspension system in the vehicle suspension case study as the vehicle passes over two potholes. Seven field runs are shown for the regions of (*a*) pothole 1 and (*b*) pothole 2. Sixty-five runs of the computer model are shown for the regions of (*c*) pothole 1 and (*d*) pothole 2. Vertical lines indicate the reference peak locations at which the curves were registered for easier comparison.

— $y^R(\mathbf{x}; t)$ is the time-history curve of resulting forces on the suspension system when the suspension system characteristics are $\mathbf{x}$, with t ranging over the time range of the vehicle test.

- **Computer model:** A finite element PDE computer model of the effect on the suspension system of the vehicle being driven over the test road was developed.
  — It depends on $\mathbf{x}$ and also on unknown calibration parameters $\mathbf{u} = (u_1, u_2)$, which are energy losses (called damping) that occur in mechanical systems.
  — $y^M(\mathbf{x}, \mathbf{u}; t)$ is the time-history force curve (at inputs $\mathbf{x}, \mathbf{u}$) resulting from running the computer model.
- **Field data:** A test vehicle was run seven times over the road containing the two potholes.
  — The nominal values of the suspension system characteristics $\mathbf{x}$ for the test vehicle are known, but the real values vary because of manufacturing variability. We denote these (unknown) real values by $\mathbf{x}^* = (x_1^*, \ldots, x_7^*)$.
  — Denote the rth field time-history curve by $y_r^F(\mathbf{x}^*; t), r = 1, \ldots, 7$.
  — The statistical model for these observations was $y_r^F(\mathbf{x}^*; t) = y^R(\mathbf{x}^*; t) + \epsilon_r(t)$, where the $\epsilon_r(t)$ are realizations of a certain error process (not discussed here).
- **Model data:** The computer model of the vehicle was run at 65 input values of $\mathbf{z} = (\mathbf{x}, \mathbf{u}) = (\mathbf{x_1}, \ldots, \mathbf{x_7}, \mathbf{u_1}, \mathbf{u_2})$; let $\mathbf{z}_r = (\mathbf{x}_r, \mathbf{u}_r)$, $r = 1, \ldots, 65$, denote the corresponding input vectors, which were chosen by a Latin hypercube design over the rectangle of uncertainty ranges for the 7 inputs. Let $y^M(\mathbf{z_r}; \mathbf{t})$ denote the rth computer model time-history curve, $r = 1, 2, \ldots, 65$.

Figure 1 shows the forces impacting one key location on the suspension system in the time interval as the vehicle passes over the two potholes.

2.2.2. The analysis and results. The analysis proceeded (we omit the details) by

- registration (aligning) of the field and model force curves (essentially scaling time so the driven car matched the computer model car as closely as possible);

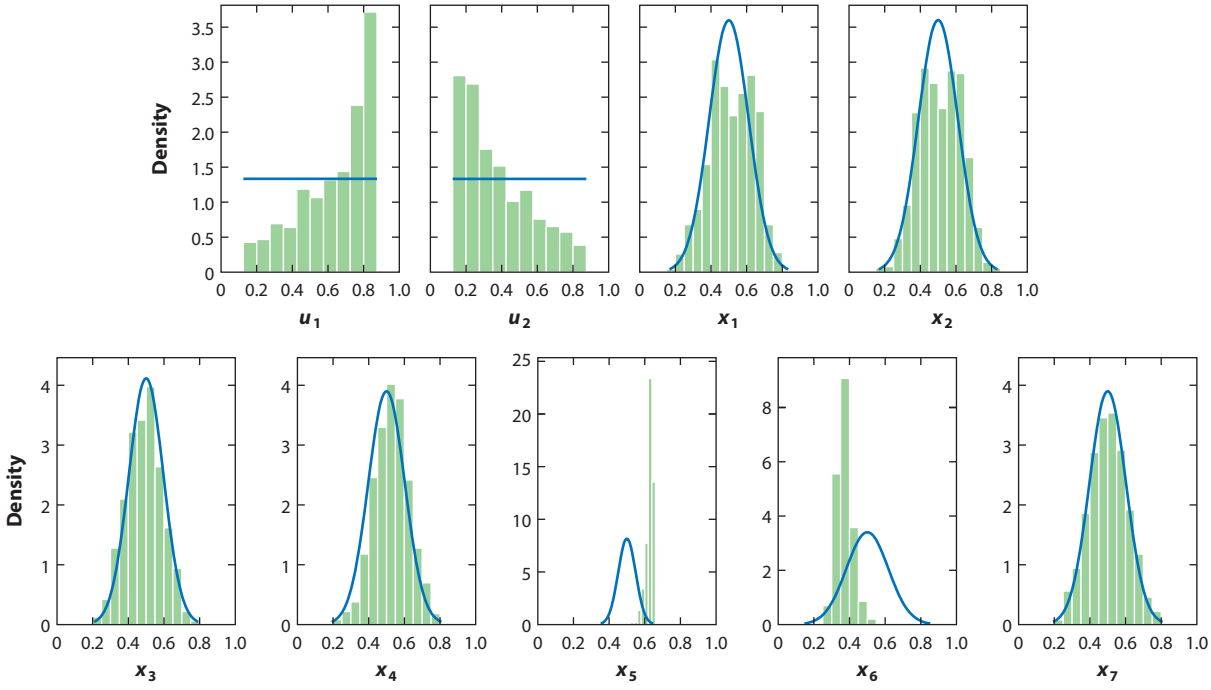

Figure 2

Posterior densities for the calibration parameters $\mathbf{u}^*$ and input parameters $\mathbf{x}^*$. The prior densities are the solid blue lines.

- employing a wavelet representation of all force curves, in which time t occurred only in the wavelet basis functions $\psi_j(t)$, and the wavelet coefficients w_j depended on the inputs (and calibration parameters for the computer model force curves);
- introducing a zero-mean Gaussian process prior to represent the model discrepancy corresponding to each wavelet coefficient;
- for computational reasons and to allow inference at new inputs, replacing each function w_j by an emulator (a Gaussian process approximation to that part of the computer model);
- assigning priors (a mix of subjective and objective) to all unknowns; and
- performing what is now called a modular Markov chain Monte Carlo (MCMC) computation (Liu et al. 2009).

This resulted in having posterior distributions for all unknowns. **Figure 2** gives the posterior densities of the calibration parameters $\mathbf{u}^*$ and the inputs $\mathbf{x}^*$; interestingly, except for x_5 and x_6, the posteriors did not differ that much from the prior densities (the solid blue lines in each figure), indicating that there was not much information in the field data about these parameters. Also, the posterior distribution of x_5 is rather suspect—shifting to the right end of its range—suggesting that the variable might be behaving as a tuning parameter to fix problems with the model. This would have been much more problematic had a discrepancy function, $b(\cdot)$, not been introduced; then, many variables might have (inappropriately) acted as tuning parameters, shifting away from their true values to (inappropriately) try to correct for model discrepancy. (See, e.g., Plumlee 2017 for ideas to address this.) More generally, note that, when estimating the parameters of a structurally perfect model, all proper scoring procedures will agree on the correct value of the parameter, which is indeed the value that generated the observations. When the model is structurally imperfect and

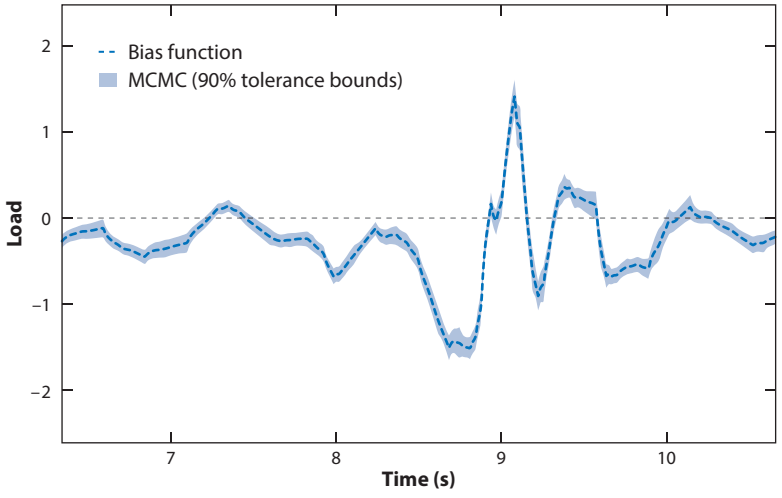

Figure 3

Posterior discrepancy curve estimate and 90% confidence bands for the first pothole region. Abbreviation: MCMC, Markov chain Monte Carlo.

no value of the parameters will yield probability forecasts consistent with the observed outcomes, then different proper scores will yield different optimal values of the parameter; arguably, the parameter in this case is not uncertain but indefinite: it is not that the estimate is imprecise but that no true value exists to be identified (Du & Smith 2012).

Figure 3 presents the posterior median and 90% confidence bands for the discrepancy function $b(\cdot)$ at one of the pothole regions. Clearly there is nonnegligible discrepancy in the computer model.

Also available is the posterior density for the real process being modeled; this estimate of reality (for the field tested vehicle) incorporates all of the information, including the field data runs, prior information on inputs, computer model runs, etc. In reality, however, only the field data significantly influence this analysis, and so the answer differs little from that obtained by standard statistical analysis of the field data alone.

Recall that the real purpose of computer modeling and the UQ analysis is to create a methodology for extrapolation to inputs $\mathbf{x}$ that were not in the domain of the inputs from the field data. One cannot typically extrapolate a statistical model outside the range of observed data, but the hope is that one can use the computer model to assist in this extrapolation. Alas, this can work only if the discrepancy function extrapolates to inputs outside of the original domain, and it is not at all obvious that this can be done.

In the present case study, this was attempted, with the results reported in **Figure 4**. The study team was presented with a different vehicle, with very different $\mathbf{x}$, for which field trials were done (the red curves in **Figure 4** are the results for the region near pothole 1), but the field data were not provided to help in prediction of the real process. Instead, the real process was predicted using the methodology above, along with 65 computer model runs (the yellow curves in **Figure 4**) for the new $\mathbf{x}$ (the 65 input values were determined by another Latin hypercube design over the input regions). The predictions of the real process are given in **Figure 4**.

The results are interesting. For the most serious force impact—just before time 8—the raw computer model runs were far from accurate, and yet, by including the discrepancy explicitly, the UQ analysis properly predicted the real data. But, for a lesser event, just beyond time 10, even with

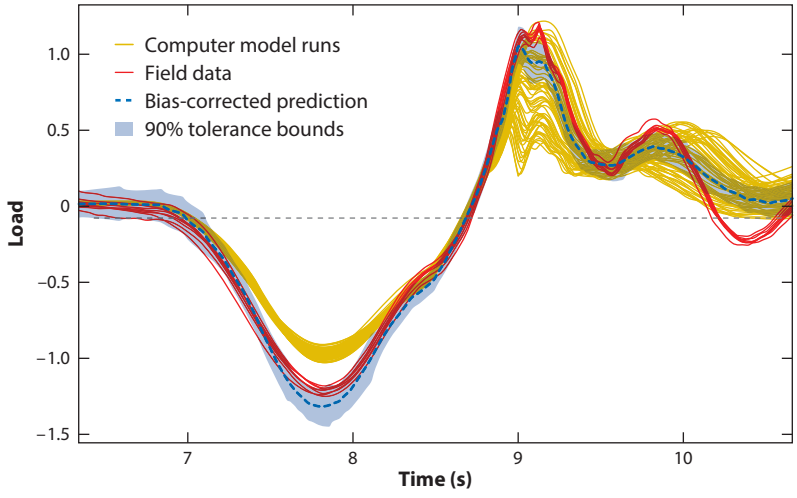

Figure 4

Prediction of the force curve in the region of pothole 1 for a new vehicle, utilizing the complete posterior distribution developed earlier, along with 65 computer model runs for the new vehicle (*yellow lines*). The red curves are the revealed (after the fact) test runs of the new vehicle, and the blue lines give the median and 90% confidence bands for the uncertainty quantification prediction.

the discrepancy correction, the UQ analysis did not predict the real data. That dip (in the real-data red curves) was never seen in any previous field or model data and so was missed. An interesting technical point is that discrepancy correction by simply adding the estimated discrepancy to, say, the posterior mean of the computer model output does not work here because of a large posterior dependence between the discrepancy and the other parameters; full posterior analysis is needed to achieve even this partial success.

2.2.3. Conclusions from the case study. What can we conclude from this case study of a UQ analysis? First, it seemed to be reasonably successful. The posterior uncertainties for the unknowns all seemed reasonable, with the possible exception of the posterior density of x_5, which may have been acting as a tuning parameter for the computer model. But, again, any other analysis (non-Bayesian or not involving a discrepancy function) would likely have suffered much more from such overtuning. And, in terms of meeting the overall goal of good extrapolation to new scenarios, the methodology mostly worked. But there are four warning signs:

- The methodology missed a correct extrapolation in the region just beyond time 10 because the behavior exhibited there had never been observed before.
- The process being studied, while complex in some sense, is almost trivial compared with processes currently being modeled (e.g., climate), so the implications of partial success here should be taken with a grain of salt. In particular, the analysis was able to pin down the discrepancy function reasonably well, a feat that is unlikely for truly complex problems.
- Related to the second point is that the unknown inputs and calibration parameters had reasonably precise prior distributions; if these were more uncertain, pinning down the discrepancy function would have been much more difficult.
- A full posterior analysis was needed for successful prediction, and this is difficult in more complex problems. Indeed, even in this problem, a modification (the modular MCMC) was needed for successful computation.

3. VARIETIES OF CASE STUDIES

3.1. A Classification of Uncertainty Quantification Problems

The modeling world is vast, and UQ is a fundamental aspect of each modeling enterprise. In this section, we introduce several additional case studies to illustrate issues that arise. We then refer back to them in discussion of the issues.

Applications of the formalism fall into at least two distinct cases. These cases are based on the context the practitioner (decision maker) faces when interpreting UQ, and they also restrict the nature of the analyses open to the practitioner.

- Weather-like tasks:
  - Similar decisions, in similar contexts, are made frequently (e.g., daily or monthly).
  - A sizeable forecast-outcome data archive is available.
  - The models involved have a long lifetime relative to the lead time of the forecast: out-of-sample additions to the forecast-outcome archive will soon be available.
- Climate-like tasks:
  - The decision is effectively one-off; the decision not to act, say, to gather more information, has potentially significant costs.
  - The forecast-outcome archive is effectively empty (the task is one of extrapolation).
  - The model has a short lifetime compared with the lead time of the forecast.

Although geophysical terms are used in these labels, the nature of the tasks generalizes: Questions of quality control in a factory with 128 identical widget-making machines are weather-like; the design of a new high-speed rail infrastructure (or the choice of which new aircraft to take into production, which America's Cup yacht design to build, etc.) is climate-like. And, of course, there are intermediate cases, seasonal-like tasks, where one has small forecast-outcome archives and moderate-lived models for which limited out-of-sample data become available before the modeler retires.

3.2. The Case Studies

To illustrate the issues that will be raised with the above UQ formalism, we will repeatedly refer to four case studies, described for convenience of reference in this section.

3.2.1. Pyroclastic flow. A computer model, TITAN2D, together with UQ, has been used to predict the risk hazard from pyroclastic flows on the Soufrière Hills volcano on the island of Montserrat (Bayarri et al. 2009a, 2015). TITAN2D is a PDE model that incorporates a digital elevation map of the island and models how the pyroclastic flow moves down the mountain, driven by gravity and topography. The key uncertain inputs to the computer model are $x_1 =$ the volume, V, of the pyroclastic flow; $x_2 =$ the direction, φ, at which the flow begins to travel; and $x_3 =$ the basal friction, b, which is the friction of the flow with the ground and is the cause for the flow to slow down and eventually stop. The light gray areas in **Figure 5** are pyroclastic flows on Montserrat. The directional influence of the flows is obvious. The influence of volume is simple: The longer flows are typically higher volume.

3.2.2. Medium-range weather and seasonal forecasting. In light of the practical challenges of performing a Bayesian analysis, an ad hoc approach of running Monte Carlo samples of historical emulators (effectively previous, simpler versions of the current weather model, these are also called predecessor-emulators) was developed in the context of weather forecasts; the first of

Figure 5

The light gray areas indicate pyroclastic flows that have happened from the top of the Soufrière Hills volcano on the island of Montserrat. Clearly, the direction of the flow and the volume of the flow (determining how far down the mountain the flow proceeds) are crucial inputs. Image adapted with permission from the National Aeronautics and Space Administration.

these EPS became operational in 1992. See Section 4.2 for discussion of the extent to which these ensembles can be interpreted probabilistically.

Figure 6 shows the component parts and results of the UK Met Office's seasonal forecasting system. The purple and yellow asterisks show the observed values of the target (average UK temperature) in each year since 1981. These values are used to estimate the climatological distribution or climatology: the distribution in the absence of any current observations (*red curve*). The blue plus signs are the model-target values from each member of an ensemble of model simulations; these are then interpreted as a probability density function for the average (August to October) 2018 temperature of the United Kingdom (*blue curve*) via kernel dressing and blending with climatology (see Bröcker & Smith 2008). The two distributions shown at the bottom of the panel indicate that August–October 2018 is likely to be warmer than a random draw from the climatology. The ability of this simple graphic to hide the depth of insight and investment that went into its making is rather astounding. Each of those model-target points utilized measurements from an observing system of oceanic, atmospheric, and space-based instruments; the model employed approximated the primitive equations of the atmosphere and the ocean, coupled, in a state space of many million dimensions. Hundreds, if not thousands, of physical parameters and model parameters were assigned values. In practice, a handful of initial states (fewer than 128) are selected in this vast state space, integrated forward in time under the model, and then used to quantify the uncertainty as a probability forecast.

Modern weather forecasting follows a similar route, although the details of the ensemble formation schemes used vary. The take-home point here is that while these schemes are often more informative than climatology (Hagedorn & Smith 2009), they are rarely, if ever, well

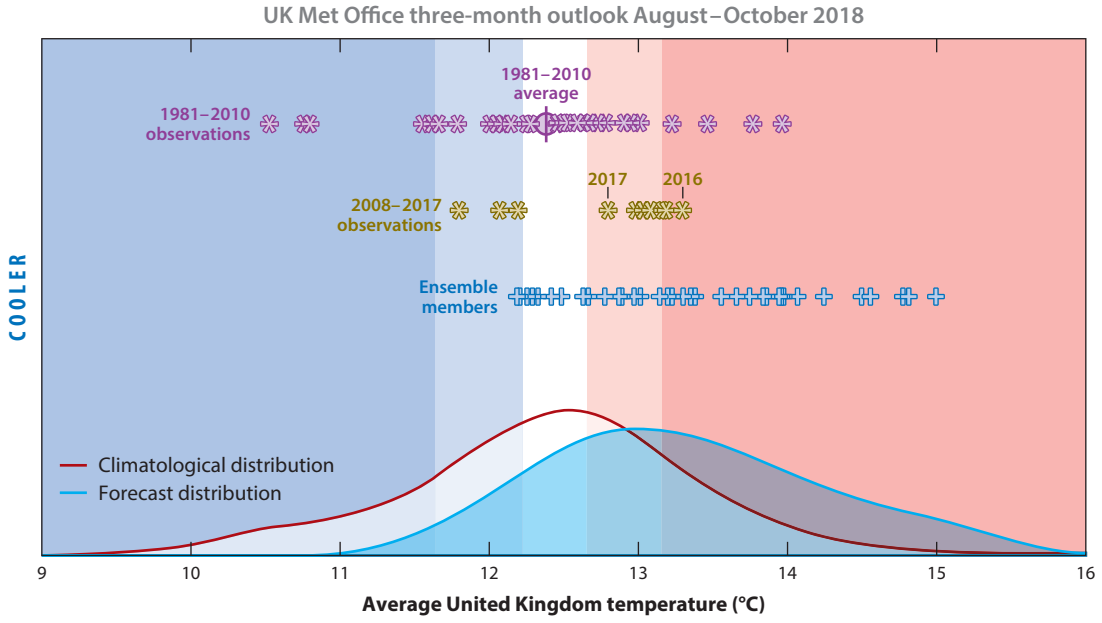

Figure 6

Probability forecast by the UK Met Office for the three month average temperature over August–September–October 2018, issued in July 2018. Adapted from the UK Met Office three-month outlook for August–October 2018 with thanks for discussion of the original image.

calibrated. The next section illustrates how they can be informative. An alternative approach is to restrict attention to questions where they are, arguably, well-enough calibrated to be useful (see Section 5.3.2).

3.2.3. Numerical weather prediction for St. Jude's Day. Numerical weather prediction proceeds along a similar path as the seasonal forecasting illustration, but with significantly more vigor; ensembles of ≈64 members are launched ≈4 times a day and integrated two weeks or more into the future. These ensemble members, simulations generated with a simpler, lower-resolution model, in some sense emulate a more complex, even higher-dimensional (hi-res) simulation that is integrated about 10 days into the future. The hi-res model is significantly more expensive computationally than each individual ensemble member, and their relative value in terms of average information content can be quantified (Hagedorn & Smith 2009).

The left panel of **Figure 7** shows real-world observations of the maximum wind speed occurring during the St. Jude storm of October 2013; the observations are made at specific points in space and spread over time. The right panel shows the weather model's version of the storm. There is a difference of type between wind, as seen through the observing system, and model-wind, as neatly defined on a space-time grid. The center panel shows a simulation at the same time corresponding to one ensemble member launched 15 days before the storm. How can we use this foresight?

3.2.4. Climate. Good weather models are, in fact, simplifications of good climate models. On a timescale of weeks or even months, many aspects of the Earth systems are effectively static (or quasi-periodic). The computational resources required to simulate these aspects are almost

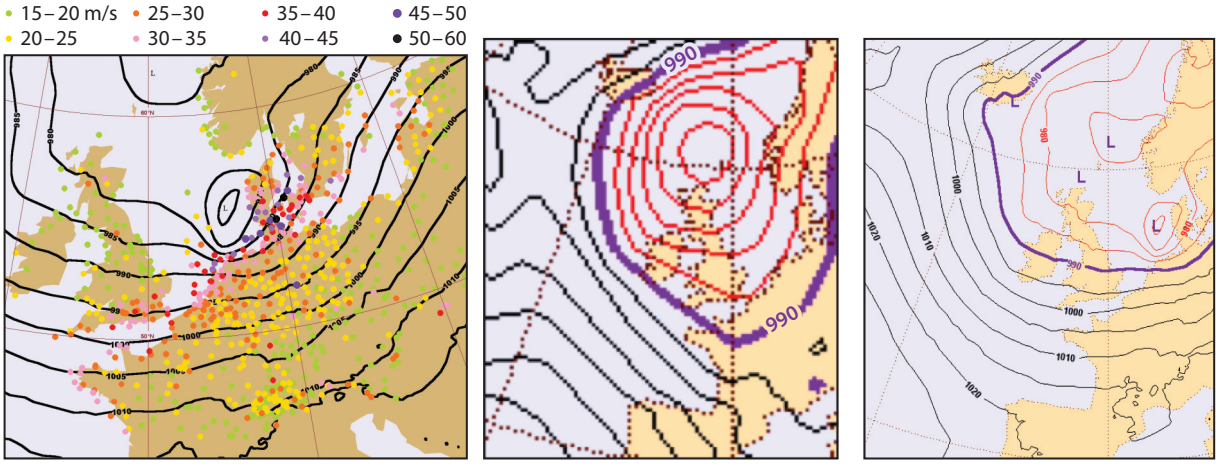

Figure 7

The St. Jude storm of October 2013. (*left*) Observed maximum wind speeds. (*center*) A member of the ensemble forecast launched 15 days before the storm. (*right*) A model-state, given the observations on the day. Images courtesy of the European Centre for Medium-Range Weather Forecasts, with thanks for its assistance in locating and providing these panels.

certainly better invested elsewhere when modeling the weather of the Earth system. Climate models, however, must include these more slowly varying aspects of the Earth system, often thought of as corresponding to ice, the deep ocean, and changes of vegetation type, for example. That said, realistic climate models must include all feedbacks that impact climate on all timescales. This includes feedbacks due to weather, since both minor, methodologically uninteresting phenomena and short-lived, high-impact phenomena can yield feedbacks that impact future climate. Such known neglecteds, which must be neglected today due to technological limitation (i.e., computing power), constrain how far into the future climate models of the day are likely to be informative; climate science can quantify this lead time.

4. ISSUES WITH THE FORMALISM

The formal UQ enterprise discussed above is a natural way to represent the problem. And, if all the probabilistic modeling that is necessary in the formalism could be done accurately and the Bayesian computation implemented, then we would not question the answer. (It could still be wrong, of course, but is defensible as a success, in that it was a mature analysis yielding probabilities that were thought to reflect the future: both adequate for the purpose and the best that could be done conditioned on the information in hand.)

To what extent can general prediction tasks be cast into the formulation above? Nine challenges to the formalism are noted in this section; the discussion includes how one might discuss the extent to which one can achieve even the limited definition of success above.

4.1. Available Data on Reality

Observations of the real process, when available, are arguably never complete, because only some of the inputs and outputs (drivers of the real process) will be known, and those that are observed

may not simply relate to the inputs and outputs of the computer model. Typical meteorological observations are made on scales of less than a centimeter and less than a second. Typical weather model-state variables are defined on discretization grids of order a kilometer and 15 minutes. In climate-like tasks there are no (or very limited) data on the actual targets. For pyroclastic flows, there is a severe lack of data on the nature of past flows, especially those that ran into the sea. In such scenarios, it does not seem possible to define and evaluate the discrepancy term in UQ, at least in terms of discrepancy of the overall process. (One might hope to do better for particular features of the model output.)

4.2. Limited Model Runs and Emulation

Models are often built because it is impossible (or too expensive) to observe the real process under study sufficiently; emulators are built when it is too expensive to run the model sufficiently. That said, a considerable number of model runs are required to construct an emulator (the more inputs and unknown parameters in the model, the more runs are needed), so a lack thereof could short-circuit the formal UQ process.

One common effort to ameliorate the problem is to develop emulators based on only the important inputs or parameters, typically fixing the others at some values. Doing so recasts the tasks back into the lower-dimensional space statisticians are more accustomed to working in, but away from the degree of complexity physical scientists deem necessary to save the phenomena. The relevant point here for UQ is that the uncertainty discarded by this truncation to important subspaces must be quantified and shown to be small. This is a critical component of uncertainty guidance, one more easily quantified than the step from the real world to model land.

Finding the important inputs or parameters and judging the effectiveness of the resulting emulator has been the subject of considerable research (Sobol 2001, Oakley & O'Hagan 2004, Linkletter et al. 2006, Schonlau & Welch 2006, Tarantola et al. 2007, Savitsky et al. 2011, Le Gratiet et al. 2014). This approach can work when the structure of the model is, in fact, more complex than required. Indeed, it can even result in a better emulator (as in statistical regression, including a variable of little importance can worsen predictive results because of the noise introduced in its estimation).

Even with comparatively simple models, construction of an accurate statistical emulator (and verifying that it is so; see Bastos & O'Hagan 2009) can be challenging. For insight into the challenge and literature, see Gu et al. (2018) and the associated R package (Gu et al. 2016b). Also, Gaussian processes are necessarily aimed at inputs that are continuous, but often the inputs are discrete; ideas for how to deal with such situations are provided by Qian et al. (2008) and Storlie et al. (2015). Constructing emulators with high-dimensional inputs and high-dimensional outputs can also be challenging, the latter less so, as indicated by Rougier (2008), Higdon et al. (2008), Paulo et al. (2012), Fricker et al. (2013), Spiller et al. (2014), and especially Gu et al. (2016a). Finally, creating emulators that try to mimic the dynamics of computer models is very challenging (see Conti & O'Hagan 2010).

Because of these difficulties, various other strategies have been utilized for UQ. One natural strategy for emulating today's state-of-the-art model is to use the state-of-the-art models of yesteryear; denote these as predecessor-emulators. The idea is to somehow use the predecessor-emulators to assess uncertainty, while using the state-of-the-art model for the base prediction. While this can work for some predictions, there are inherent limitations when the predecessor-emulators lack the complexity of the current state-of-the-art model, much less that of the true process. Consider two examples where emergent processes are both observed and understood. The first is hurricanes, which simply do not form in low-resolution models of the atmosphere.

The second is orographically driven rainfall due to mountains. Again, the predecessor-emulators have systematically more slowly changing orography: In the model, mountain ranges will not reach the height of the real mountains that they represent, sudden shifts in elevation (plateaus) will rise more gradually in the model, and so on. The disconnect between the model and predecessor-emulators on targets like flash flooding precludes their effective use in uncertainty analysis. The search for statistical emulators of these targets on weather forecasting timescales has been fruitless for the past half-century. On seasonal timescales, statistical models (called empirical models; see van den Dool 2007) are often competitive with today's best simulation models; they do not appear to emulate them.

Another approach to developing an emulator is the ensemble approach: Choose a range of values of the inputs and parameters that is judged reasonable, and run the (state-of-the-art) model over a smallish selection of values over this range. Then, treat the resulting runs as a probabilistic sample of the real process. This is the basis of UQ in weather forecasting and is also utilized for UQ in climate modeling (see Stainforth et al. 2005 and interesting subsequent discussions regarding interpretation of new effects seen when an ensemble size is increased by two orders of magnitude, in later articles that criticize the Stainforth et al. study). If one knew the prior probability distribution over inputs and parameters, and if the model had no discrepancy from reality, one would, indeed, have what could be considered to be a sample from the true process, an accountable ensemble (see Smith 1995). In both weather forecasting and climate modeling, however, the prior probability distribution is unknown, the initial input state space is enormous ($\approx 10^7$), there are hundreds of unknown (arguably ill-defined) model parameters, and there is structural model error. Furthermore, a vast number of ensemble members would be needed for an accurate UQ analysis, and it is impossible to obtain a large number from state-of-the-art weather or climate models. The information from ensembles has proven value in weather forecasting and represents a major step forward from the previous practice of just running the model once at fixed inputs and trusting/interpreting the result in both weather and climate; nevertheless, it is difficult to know how to interpret these ensembles quantitatively in UQ, particularly in terms of uncertainty guidance.

4.3. The Real World Versus Model Land

Suppose we are forecasting the temperature at a point on the surface of the Earth. The model-state x consists of a grid of points upon which a PDE is being integrated numerically, and arguably, the true x is an evolving three-dimensional field. One of the observed components of x is a thermometer reading at 06:32. The representational error describes the shortcoming of taking the thermometer reading to represent the grid-point value of x. It is a shortcoming of the model. In the same way, there are shortcomings of interpretational error, presuming that the model-future grid-point x corresponds to a future thermometer reading in the same vicinity.

The traditional framework assumes not only that the real process/system admits a nontrivial mathematical description (we will assume it does, but note the discussion of the Russell Map in Smith 1997) but also that the functional structure of the system and its state space is identical to those of the models. This is often known not to be the case. Consider, for example, cases where the system input is thought to be a continuous field evolving smoothly in continuous time; the model input of a computer model will be discrete in space, time, and value (due to digitization). In this case, $\mathbf{x}$ of the system and $\mathbf{x}$ of the model are different types of entities, and the notion of distance operations like subtraction (see Lorenz 1985 and references therein) is significantly more subtle than in the traditional formalism. The question is the extent to which this matters in a given task.

This issue also extends to the definition and interpretation of model discrepancy. The traditional formalism takes the discrepancy to be a function of model-state **x** and, in practice, this can fail in two ways. First, the model-state typically does not define the system-state; this case of model inadequacy is discussed by Judd & Smith (2004), where it is called the missing subspace. Second, the full dependence on **x**, even model-**x**, is often suppressed for convenience. In climate model studies, for instance, variability due to small changes (at machine precision) in initial state may be significant. Indeed, the discrepancy can appear fractal (far from smooth) at machine precision. How is the uncertainty introduced by this fact to be accounted for? If changing the compiler setting (much less the computer hardware) leads to a different posterior distribution, the study is known to be flawed. The situation is similar to that in which the final distributions in a stochastic model vary with the particular seed of the pseudo-random number generator; we can find ourselves in the position of knowing the study is incomplete (at best, that the current sample distribution simply has not converged) without being able to test further for the reasons why.

A final point: In nonlinear models, small uncertainties can grow rather quickly. In the statistical formulation, the state variables of the model correspond effectively to observations. The observing stations shown in the left panel of **Figure 7** are of a different type from the model-state variables of a weather model. How might UQ track the uncertainty due to this aspect of data assimilation? And, of course, there is the simpler problem of the sheer dimensionality of the space of the hi-res model: How can we describe, much less track in time, the diversity of paths with finite initial spread in each of several million components of the state of a highly nonlinear system? The task can be formulated clearly within the Bayesian UQ framework; the question is how to include (guidance regarding) the uncertainties that arise when attempting to deploy the framework in practice. UQ must include guidance on uncertainty due to this real world–model land disconnect.

Statistical models often work in a model-state space consisting of the observed quantities themselves. This eases vastly both data assimilation (going from the space of observation to the model-state space) and also model interpretation (returning from the model-state space to the real world).

4.4. Issues With Model Discrepancy

Without question, the presence of discrepancy in the UQ formalism is a significant extension, in terms of our ability to formulate a question statistically; less clear is the extent to which we can solve the problem as formulated today, or the extent to which simplifications deployed today reflect a solution to the UQ problem that introducing the discrepancy allowed us to formulate properly. A high-dimensional discrepancy or a lack of smoothness can hamstring its deployment.

4.4.1. When the model is flawed. Modern data assimilation techniques yield information on state-dependent model error in dynamical models. Discrepancy provides a novel approach to this task. Perhaps the most important use of model discrepancy is identifying flaws in the model, with the hope that they can be corrected. In the pyroclastic flow case study, a model flaw was observed with the basal friction parameter. Since the pyroclastic flows being studied consisted primarily of silicon, the friction value 15 (that for silicon) was used initially. When it became clear that the probabilistic predictions of risk with the model were simply wrong (the computed 10-year risks of certain events were very low, even though those events were frequently observed in the relevant decade), an examination of the model ensued, with the friction value of 15 standing out as the prime suspect.

Subsequent study revealed that massive pyroclastic flows simply behave differently than ordinary silicon. By utilizing other observed pyroclastic flows (for which both the volume and the length of the runout of the flow were accurately known), the basal frictions of the flows were

better estimated. A new posterior distribution of basal friction, conditional on volume, was determined (Ogburn et al. 2016), and by using an emulator, the resulting predictions appeared to be very sensible.

4.4.2. The big surprise. Big surprises occur when the target system behaves in a way no current model can capture correctly, even in retrospect. Some impacts, say the probability that the Earth is hit by a large asteroid, have been neglected in climate predictions. In this case, the model-based distributions would prove misinformative: The fact that the probability of this big surprise is very small is informative. What is the relevant dominant uncertainty (say, due to known neglecteds) in climate predictions for 2100 (Smith & Petersen 2015)? Bank of England probability forecasts noted explicitly (in the captions) that the predictions assumed Greece did not leave the Eurozone. This arguably counts as noting a big surprise; the Bank did not give alternative forecasts conditioned on Greece leaving. UQ is incomplete without a quantitative estimate of the irrelevance of model-based probabilities, and incomplete UQ is actively misleading when that probability is unstated and not vanishingly small.

4.4.3. Known neglecteds. Building physical simulation models often requires omitting processes known to be important in order to achieve a viable model. In the pyroclastic flow example, TITAN2D employs a two-dimensional approximation to the three-dimensional truth. Mountain ridges are not well represented in any 2018 climate model due to issues of numerical resolution; as a result, precipitation due to orographic effects is unrealistic. Clouds are not computed explicitly even in weather models but are instead accounted for in ad hoc fashions. In simulations for decision support, such known neglecteds, along with more foundational uncertainties in the Earth system, make a formal UQ regarding the climate one of the most pressing challenges of the day. Known neglecteds are dealt with either by some ad hoc process or by simply ignoring them on the grounds that they do not matter much on the prediction lead times of interest. While ignoring known neglecteds willy-nilly may be good practice in pure research, a predictive context requires quantification of the uncertainty introduced by these scientific shortcuts; clarity here is of particular importance when omitting a known neglected is required by technological constraints.

Model developers should quantify their expectation that the simulation will fail to reflect reality because of these known neglecteds, by providing a probability that they will lead to a big surprise for a given target, as a function of lead time. In weather-like tasks, where we have a large forecast-outcome archive, statistical models of model failure can also be generated from the archive; these are useful in themselves and can also serve to better inform experts concerning the probability of a big surprise.

Similar issues can occur in purely statistical models. A financial model, for example, may assume there is always sufficient liquidity in the market. An econometric model may presume rational behavior by people, when it is known that behavior is frequently not rational. Most missing data models assume data are missing at random, when this is frequently questionable. Nevertheless, the situation with known neglecteds in state-of-the-art large-scale computer modeling is typically more severe than in statistics, because there is usually nothing that can be done about it today other than reducing the forecast lead time (in statistical modeling and empirical modeling, the model can be elaborated if needed).

4.5. Prior Distributions

Even when the model inputs or model unknowns are of relatively low dimension (as in the vehicle suspension and pyroclastic flow examples), assessing prior distributions for the unknowns can be

challenging. This was dramatically illustrated with the pyroclastic flow example, where initial prior assessment of basal friction was seriously flawed, and an extensive side study was needed to develop a reasonable prior for basal friction.

When there are hundreds or thousands of unknowns, many of which may have amazingly complex interdependencies, the problem of specification of the prior becomes extremely challenging; values of $\mathbf{x}$ may, with nonvanishing probability, lie on low(er)-dimensional manifolds or fractal sets. Furthermore, the distribution of ϵ may be state dependent, even in the case of additive observational noise.

4.6. Extrapolation

Even if the statistical formalism can be implemented, there can be no assurance that the simulations will reflect reality in extrapolative settings. (This is, in large part, simply Hume's Problem, also known as the problem of induction; Howson 2000.) The discrepancy correction cannot, in general, be trusted outside of the range of the real data that were used to determine the discrepancy. Scientific reflection on plausible impacts of known neglecteds can provide an envelope outside which the model is likely to be misinformative; no method to indicate where the model is not misinformative is known. In the vehicle suspension system example from Section 2, a successful extrapolation was performed, but this might have just been luck (and the extrapolation was not perfect). To some extent, every weather forecast is an extrapolation into the future conditioned on the belief that today's understanding of climate and weather will also hold tomorrow and the hope that we have turned extrapolation in time into interpolation in model-state space. In climate-like cases, of course, where the boundary conditions have never been observed and the model cannot have been tuned, this assumption has little support.

The situation, in terms of extrapolation, with the different climate models is interesting to ponder. The climate models do not at all agree as to the current climate, but they roughly agree as to the change in climate that will be happening. So the models have revealed themselves to be flawed (at least some of them), but can we trust their rough common extrapolation (Parker 2011)?

5. ISSUES WITH PROBABILITY

Several of the concerns that have been raised with the statistical formalism revolve around the capability of probability as the vehicle to express uncertainty. Our view is that probability is the correct vehicle (in the sense that any other vehicle is inferior) but only when properly utilized; this last phrase carries with it a host of philosophical and practical issues that are explored in this section. Perhaps the most poignant issue is how to proceed when we have generated probabilities but do not believe all the assumptions under which they were derived; we save discussion of this for last.

5.1. Epistemic Versus Aleatoric Probability

Aleatoric probability is probability that arises due to some known random mechanism. Epistemic probability is probability used to describe uncertainty about some fixed, but uncertain, quantity. In UQ, aleatoric probability arises from many sources, for instance, randomness in field data and randomness in model inputs [e.g., in the pyroclastic flow example, the inputs $\mathbf{x}$ = (flow volume, flow direction) are random initial conditions]. Epistemic probability would be used to represent uncertainty in quantities like calibration parameters (e.g., the damping parameters $\mathbf{u}$ in the vehicle suspension example) and model discrepancy.

While nearly everyone would agree (at least in the abstract) with the use of probability to deal with random mechanisms, many are uneasy with using probability to deal with epistemic uncertainty. For instance, in the nuclear regulatory world, there are many fixed but unknown parameters that are constrained to lie in intervals; a Bayesian would typically assess a probability distribution over the intervals, whereas the nuclear regulators do a worst-case analysis based on finding the variation of the quantity of interest as the unknown parameters vary over the intervals. Others accept the use of probability to deal with epistemic uncertainty but reject the common mechanisms used for implementation. The challenge here is that something we do not know or have knowingly neglected has a major impact on the outcome we are predicting.

5.1.1. Misunderstandings from not using the complete probability distribution. To begin, we illustrate one common type of misunderstanding that arises due to use of too-simplistic probabilistic thinking. Consider the following two scenarios involving missile production:

- Scenario 1: A production process for missiles has a 10% random failure rate. So a given missile has an (aleatoric) probability of 0.1 of failing.
- Scenario 2: There is a 10% chance that the design of the production process was flawed, in which case all the missiles will fail. A given missile still has an (epistemic) probability of 0.1 of failing.

We have often heard the argument that, while the probabilities are both 0.1, these are two very different situations that require very different handling, and hence, this is viewed as a clear example of the need to think differently about the two types of probability. The more relevant probabilistic analysis, however, would be to look at the joint distribution of missile failure. In Scenario 1, each missile has an independent probability of 0.1 of failing. In Scenario 2, the joint probability distribution is that they all fail with probability 0.1 and all work with probability 0.9. These are two very different probability distributions. The two situations also differ in terms of learning; there is nothing to be learned in Scenario 1, while testing one missile in Scenario 2 will reveal the situation.

So whether one has epistemic or aleatoric uncertainty affects how one develops the joint probability model and can affect learning about the probability model from data, but probability is capable of handling either type of uncertainty well. We conclude the missile discussion by asking the reader to cast themself as the decision maker in a tactical situation, faced with using one missile or not, with no possibility of using others. Would you make a different decision in the two scenarios, or simply proceed by thinking that the missile failure probability is 0.1?

This is not to say that there are no differences between aleatoric and epistemic probabilities; the nuclear regulator situation is a prime example. We are just saying that differences are not about probability itself, but about the intended use of probability.

5.1.2. Mixing aleatoric and epistemic uncertainties. Disturbing convolutions of aleatoric and epistemic uncertainties can occur, especially in the case of nonlinear models. It is often the case that, on general principles, we know the dynamics of a system will be confined to a manifold (or perhaps a fractal attractor) of dimension less than that of the state space. Our epistemic insight need not be constructive: We know the system evolves on a manifold, but we do not know any details regarding the manifold. Aleatoric uncertainties in the observations still cloud the initial conditions and the outcomes, but our epistemic insight means that uncertainty in the initial condition cannot be so simply accounted for, as in the case where one might believe in a uniform prior distribution of initial conditions in this region of the state space. This complication is easily dealt with in principle, but in practice it is unclear how to quantify the initial uncertainties. And

in practice, of course, the model is not structurally perfect, and one can expect a loss of structural stability (Hirsch & Smale 1974); without topological conjugacy, even if one manages to construct the relevant manifold (or attractor) of the model, one can expect that this manifold will not reflect that of the system (if such a thing exists). We are not saying that probability is an inappropriate vehicle for describing this mix of uncertainty, but rather that our access to that vehicle is unavailable (Smith 2002).

5.2. Uncertain Probabilities

The purpose of probabilities is to describe uncertainty but, unfortunately, probabilities are themselves often uncertain. In this section, we discuss this issue and some of the solutions to the problem that have been proposed.

5.2.1. Interval analysis. The Bayesian paradigm is typically phrased in terms of precise probabilities, e.g., the probability of rain tomorrow is 0.4 (i.e., 0.4000000000 . . .). This is not realistic. If a weather forecaster says 0.4, they probably really mean something like "the probability of rain is between 0.35 and 0.45." If asked for a more precise prediction, they would likely be unable to provide an additional digit, e.g., 0.42. So one should rationally think in terms of eliciting interval probabilities.

The same is true for unknown input or calibration parameters. It is typical to only be able to restrict them to lying in a certain interval (or more complex joint surface). In **Figure 2**, for example, the seven parameters were initially restricted to be in certain intervals. There have been many suggestions for dealing with such intervals of probabilities, but here we discuss only three:

- Pointwise range analysis: For each vector of unknowns arising from choosing points in each interval, determine the answer, and then look at the range of such answers.
- Probability: Assign probability distributions to the intervals (possibly joint). This was done in the vehicle suspension example (see **Figure 2**); the normal distributions assigned to the first five inputs are noncontroversial, corresponding to known manufacturing variation. The uniform distributions assigned to the two damping parameters are more controversial, although assigning uniform distributions to the intervals is standard practice in some fields (e.g., high-energy physics).
- Robust Bayesian analysis: Form classes of probability distributions over the intervals, and then find the range of answers corresponding to varying priors.

To illustrate these possibilities and provide a background for discussion, we consider the following pedagogical example.

Example. A system contains m independent components, with component i having probability p_i of working over the time interval of interest. The system works only if all components work, so the probability that the system works is $P = \prod_{i=1}^{m} p_i$.

Initially, it is possible to restrict the p_i to intervals, namely $p_i \in (a_i, b_i)$, $i = 1, \ldots, m$, with the a_i and b_i specified.

Examples of the above interval analyses follow.

- Pointwise range analysis: Clearly $P \in (\prod_i a_i, \prod_i b_i)$, i.e., the lower and upper endpoints of this interval are the worst-case and best-case scenarios. This conclusion is certainly a true statement, but unless m is very small, the range will typically be useless, resulting in a statement such as $P \in (0.4, 0.99)$.

- Probability: Consider the (default) assignment of uniform distributions to each interval. Then compute the density of P using probability theory, and report, say, a 99% confidence set for p. This will give a much smaller interval for P than using pointwise range analysis, e.g., $P \in (0.93, 0.98)$.
- Robust Bayesian analysis: Form classes of probability distributions over the intervals, and then find the range of answers corresponding to varying priors. In forming the class of priors, typically, reasonable assumptions would be that values near the midpoints are often more likely than the endpoints and beliefs are typically symmetric and unimodal in the intervals. The idea would be to then consider all the probability distributions compatible with these beliefs and find the range of Bayesian answers over this class (which we return to more formally in the next subsection). Interestingly, it can be shown that the extremal 99% confidence interval over this class of probabilities happens to be exactly the same as that arising from the uniform priors.

What can we conclude? First, as stated above, the pointwise range analysis is logically sound, and if the resulting answer is useful, then fine. Unfortunately, in the vast majority of UQ problems, the pointwise range analysis results in a stated range that is much too wide to be useful.

Assigning probability distributions to ranges is not controversial when they are readily available, as in the vehicle suspension example for the first five inputs. The common default option of assigning uniform distributions to the intervals can certainly be questioned, but it gives the same answers as the robust Bayesian analysis mentioned above, which seems quite believable. But there are complications. For instance, often the p_i are dependent, which would have to be incorporated into the overall class $\mathcal{P}$ of possible priors, and this is challenging to do in a robust Bayesian way. Also, this clearly depends on m. If $m = 1$ and it is an epistemic uncertainty, it is at best hard to justify using a uniform distribution on the interval. But for large m, it seems that some kind of law of averages might come into play (averaging over small true values for some intervals and large true values for others), making the uniform probabilities reasonable.

5.2.2. Classes of probabilities. The most common objection to use of probability is the perception that it is restricted to one choice of probability distribution. This perception is reinforced by many of the axiomatic approaches to uncertainty, which state that a unique assessed probability distribution is needed to avoid problems like sure loss in betting.

This is not reality: Typically probabilities and probability distributions are themselves quite uncertain, and there exist many efforts to deal with this uncertainty. Arguably the most useful formal approach is to attempt specification of the class of probability distributions that is compatible with beliefs or scientific understanding (such a class is often called a credal set; see Levi 1983) and then study the range of answers that follow from consideration of the class. This approach has many names, one of them being global robust Bayesian analysis (Insua & Ruggeri 2012). There is no space herein to discuss this in depth, but it is useful to consider two examples. The first indicates how the big surprise can be addressed probabilistically, and the second involves combining sources of information.

Example. Suppose application of the statistical formalism leads to a predictive probability distribution $p_0(y)$ for reality y. But we assess that there is a 20% chance of the big surprise, with no idea as to what the surprise will be. This can be represented by the class of probability distributions

$$\mathcal{P} = \{0.2q(y) + 0.8p_0(y);\ q(y) \text{ being any distribution}\}.$$

A strict Bayesian might argue that one needs to assess a prior distribution over $\mathcal{P}$, which is then effectively the same as simply having a single probability distribution. But there is also the option of exploring options over the range of distributions in $\mathcal{P}$. In a decision problem, for instance, one might find that a certain decision is fine for all $p \in \mathcal{P}$, precluding the need to attempt to assess a distribution over $\mathcal{P}$.

Example. Suppose the model-based probability forecast for weather is P_{M}, and we also have available $P_{\mathrm{clim}}(x)$, the weather probability based only on historical data. We might then reasonably say we are dealing with the class of probability distributions

$$\mathcal{P} = \{\alpha P_{\mathrm{M}}(x) + (1-\alpha)P_{\mathrm{clim}}(x);\ \alpha \in A,\ P_{\mathrm{clim}} \in \mathcal{P}^*\},$$

where α reflects the trust we put in the model versus the historical data.

In weather-like tasks, historical observations allow the estimation of P_{clim} (see **Figure 6**) and α can be tuned for good performance using the forecast-outcome archive. Whether we choose to use empirical Bayes (estimate α and P_{clim}) or hierarchical Bayes (give them distributions), we expect a reasonable outcome. In climate-like tasks, however, we (*a*) have no archive from which to estimate α and (*b*) lack an experience-based source for the distribution that plays the role of P_{clim}, as we have only experiences of the climate(s) of the past (which we do not view as exchangeable with current climate). So while the credal set approach is conceptually fine, our inability to effectively determine A and $\mathcal{P}^*$ creates a dead end for quantitative prediction. Nevertheless, it clarifies the challenge, perhaps reducing our overconfidence in our current estimates of P and redirecting our attention to more policy-relevant tasks (Smith & Stern 2011, Parker & Risbey 2015).

5.2.3. Upper and lower odds. Working with credal sets of probability distributions, while logically sound, is difficult in terms of communication; thus, one typically reports just the extreme answers (for the quantity of interest) resulting from variation over the credal set.

Since odds of an event are one of the most commonly used quantities of interest, it is interesting to separately discuss reporting upper and lower odds resulting from credal sets. For instance, in the second example above, if $x > 0$ means global warming, the upper and lower probabilities for global warming are

$$\left(\inf_{\{\alpha,\mathcal{P}^*\}} [\alpha P_{\mathrm{M}}(x>0) + (1-\alpha)P_{\mathrm{clim}}(x>0)],\ \sup_{\{\alpha,\mathcal{P}^*\}} [\alpha P_{\mathrm{M}}(x>0) + (1-\alpha)P_{\mathrm{clim}}(x>0)] \right),$$

from which the upper and lower odds are readily available.

Many methods have been proposed for developing upper and lower odds, but because of our adherence to probability (as much as possible), we strongly prefer developing upper and lower odds through credal sets of probabilities. But given the complexities of UQ problems, we are not averse to specification of upper and lower odds through expert judgement, combined with the available probabilistic evidence.

5.3. Probabilities, Betting, and Decision Making Given UQ

5.3.1. Well-calibrated probabilities? Probabilities are routinely reported with the implication that they are well-calibrated, which leads to statements like "odds of 10 to 1 that it rains tomorrow imply that the odds-giver is willing to bet proportionally on either side of the proposition." But probabilities are not always well calibrated, and treating them as such opens one to the possibility of rapid ruin (see the next section).

In weather-like contexts, where the same question is being addressed repeatedly in a mechanical manner (e.g., the probability that rain will be recorded at Heathrow Airport each day), being well calibrated is computable, and almost everyone would agree that it would be wrong to report probabilities, in the context of decision-support, that were not well-calibrated. In other geophysics contexts, well-calibrated probabilities are only rarely, if ever, encountered.

UQ analyses with a discrepancy function (such as the suspension system example) have a shot at being well-calibrated. Those without (such as the pyroclastic flow or the St. Jude storm) are in limbo. Issues such as the big surprise continually stress the idea of being well calibrated. The actions of some reputable scientific organizations are difficult to interpret. For example, the provision of probability distributions for the distant future (the end of this century) at high resolution in space (25 km) and time (hottest/wettest day) appear to be inconsistent with general statements of uncertainty by the Intergovernmental Panel on Climate Change (IPCC) regarding our knowledge of global mean temperature. Claims to provide well-calibrated probabilities given only on simulations under (a set of differing) model extrapolations and expert judgement, with little substantiating data on reality, are unsettling.

Utilization of credal sets of probabilities of upper and lower odds, or reporting odds-on and odds-against based on empirical judgement or expert opinion, softens the impact of a lack of calibration. Providing the probability of a big surprise explicitly enables practitioners to more easily discard model-based probabilities that are indeed irrelevant to the task they face.

5.3.2. Example: forecast direction error. Regulators in the United Kingdom require energy companies to hold a minimum amount of natural gas based on the hi-res (point) weather reference forecast for the next few weeks; when low temperatures are forecast, the company must hold more natural gas. Suppose a gas trader is considering the supply of natural gas for two weeks from today and that buying, when the (public) forecast drops, and then selling, when the forecast rises, carry a significant cost. An ensemble forecast, interpreted (Bröcker & Smith 2008) to be a probability distribution, is also available; this is the best available probability forecast in that it has the best out-of-sample logarithmic skill. It is not, however, well-calibrated.

The trader wants to know whether or not future reference forecasts of gas consumption for the target date will move more than a distance δ from the current value. Accepting that forecasts must be probabilistic, the trader is happy to accept a yes/no answer to the question "Will the reference forecast change by more than δ with probability θ or greater?" If one interprets the forecast distribution as a probability distribution, this question can easily be answered for any pair of values (δ, θ), along with an indication as to whether the future forecast for the target date is more likely to go up, or go down, or show unusually large volatility. Cross-validation makes it clear that the answer provided by this approach is not reliable. Offering bets with odds determined by θ on those three outcomes (above, within, or below the δ-range) has a vastly higher risk of ruin than expected from the probability distribution, and the probability distribution is not well calibrated in general.

For particular values of δ, θ, however, one can sometimes develop probabilities that, under cross-validation, are arguably well-calibrated probabilities (Smith 2016). In practice, when a model is faced with inputs that are impossible given the assumptions it was constructed under, it can communicate that any probability outputs should be ignored. In studies like that reported by Smith (2016), this situation was communicated via a purple light signal, in place of the more common red, blue, and green lights. Deployment of this general approach requires ongoing monitoring. Indeed, such monitoring is a crucial component of UQ in quantitative decision support for weather-like tasks.

5.4. Combining Model Probabilities

What if we are given probability distributions, but we know they are not individually well calibrated? It is tempting to combine them in some way. For instance, suppose we are given probabilistic predictions from seven different climate models. We can certainly attempt to average them in some way, ideally weighted by the skill in prediction that each model evidences (see, e.g., Smith et al. 2009). But how to do so is very unclear, especially when they each have low, in-sample skill and/or were not each constructed in isolation. Indeed, this is essentially the problem of combining expert opinions. What one does depends very much on the degree of dependence between the experts (and climate models have very high dependence, as even independently constructed models will share some of the same weaknesses). The veracity of the model-based outcome, however, can be captured in the expert's probability of a big surprise.

There is a principled approach, which is to be Bayesian, simply viewing the given probability distributions as data (Berger 2013). But this is rarely implementable, as it requires the Bayesian to model both the dependencies of the experts (maybe possible) and the relationship of their probabilistic predictions to reality (rarely possible). This approach may prove more applicable in weather-like tasks, but the subtleties of interdependence of the probabilistic predictions due to the interrelationships of the models used leave difficult challenges.

6. MOVING FORWARD

It is critical that the uncertainty in predictions from modeling be kept at the forefront of policy and decision-making discussions. We are not questioning the enterprise of UQ; it is essential. Indeed, even though UQ has grown enormously over the past decade, we feel that it is still vastly underutilized; there is still a tendency in many areas (e.g., climate) to devote almost all resources to model development and simulations rather than assigning a significant portion to UQ.

There is general agreement that UQ in practice can deliver information on the sensitivity of a model to variations in initial model-state variables and the values of model parameters, to stochastic elements within the model, and so on. Such qualitative types of information were the initial staple of UQ; it is only more recently that statistical UQ morphed into a system claiming predictive power in the context of large simulation models.

There is also widespread agreement that UQ is most naturally performed within a probabilistic (Bayesian) framework, agreement that holds as long as there is no suggestion that the framework can be realized in any particular case. Much of UQ research is involved in extending the cases in which this framework can be applied, and the progress is dramatic; at the same time, the modeling world is developing ever more complex models for which application of the framework is increasingly difficult.

The criticisms of UQ that have been presented herein are in regards to its current implementation; it is not uncommon to see presentation of outputs from incomplete or ill-designed UQ studies as if they were decision-relevant probabilities representing the uncertainties. Incomplete UQ can lead to overconfidence in numerate practitioners, who both trust and know how to use mature probabilities.

So how do we move forward with a more sensible UQ? First and foremost, with any UQ report, it is important to clearly list which uncertainties have been dealt with in a believable fashion and which could not be addressed (e.g., model discrepancy could not be determined because of a lack of relevant field data for the process). When there is uncertainty in the probability distribution, utilization of credal sets of distributions can be a significant step forward. Discussion, in the UQ report, of the probability of a big surprise would also be welcome; by definition, this introduces significant expert judgement similar to that involved in designing the model and selecting the known neglecteds.

In weather-like UQ tasks, continual monitoring of the UQ probabilities to ensure that they are acceptably calibrated is only prudent. Even then, however, one cannot be certain of the calibration; an example is the probabilities of extreme weather events, where scarcity of data and a changing environment may well make calibration unachievable even if a lack of topological conjugacy is overcome.

It is particularly difficult to gain traction for UQ in climate-like tasks; progress here would be valuable. One possibility for getting a sense of the uncertainty would be to build different models of the process in isolation (e.g., do not allow any interaction among the simulation modelers beyond access to the observations). Would the models' long range projections converge or diverge over the next 10 or 20 years? The models could, of course, converge and still share common epistemic flaws. But when they have not converged, we gain more insight into uncertainty than when they actively work toward convergence through intercomparison. Current climate models disagree significantly on basic things, such as global mean temperature. Their biases are currently obscured by presenting changes in a value (so-called anomalies that originally brought the two coauthors together), and errors are misleadingly reported relative to the error of the best model.

We are emphasizing in this conclusion a component of UQ that could be called uncertainty guidance: the clarification of how quantitative outputs from UQ are to be used in practice. Maintaining credibility requires clarifying the utility of every probability statement in decision making. There are good examples of this: The IPCC Fifth Assessment Report, for example, states that there is a significant probability that global mean temperature in 2100 will fall outside the range of predictions of current climate models (Stocker et al. 2013). Uncertainty guidance such as this needs to become the norm. Indeed, it defines a new frontier for UQ, a frontier still rooted in probabilistic and decision-theoretic analysis.

SUMMARY POINTS

1. UQ is a critical aspect of both statistical modeling and simulation modeling.
2. The notion of model inadequacy/discrepancy (recognizing that the model is wrong) in simulation modeling is a huge step forward.
3. The Bayesian formalism of UQ is appropriate when implementable, and, in that case, no other formalism is better.
4. In practice, probabilities are often presented without the full formalism being implemented (e.g., model discrepancy was not considered), and these probabilities are not the decision-relevant probabilities one would want. Generalizations (such as the use of credal sets of probabilities or the magnitude of the probability of a big surprise) can significantly extend the operational relevance of Bayesian UQ.
5. UQ should incorporate uncertainty guidance: Communicating the limitations of each and every stated probability is a critical component of UQ.

FUTURE ISSUES

1. UQ must continue to accelerate its rapid development; recognition of its centrality to computer modeling needs to be further accepted.

2. UQ was initially highly resisted by many modelers, who simply stated that they were doing the best modeling possible. Similarly, many UQ people today resist uncertainty guidance, stating that they are doing the best uncertainty analysis possible. Some mechanism to encourage adoption of uncertainty guidance (e.g., making it necessary for journal publication) would be nice.
3. The efficacy of high-dimensional simulation, initially resisted by many statisticians who simply stated that there must be a low-dimensional model, needs to be accepted when (or only when) the high-dimensional simulations are demonstrably informative.
4. Methods for dealing with uncertain probabilities, either formally (e.g., credal sets of distributions) and/or via expert opinion (e.g., the probability of the big surprise), need to be better understood and widely utilized in UQ.
5. The role of expert opinion in the UQ process (as in almost every other situation in the world) needs to be better understood and made more quantitative where possible.

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

J.O.B. was supported by the US National Science Foundation, grants DMS-1228317, EAR-1331353, DMS-1407775, and EAR 1521855. L.A.S. was supported by EU Horizon 2020 grant ECOPOTENTIAL and the UK EPSRD Network grant CRUISSE; he is grateful for the continuing support of Pembroke College, Oxford. L.A.S. acknowledges financial support under the UK Natural Environment Research Council under grant IRIS NE/R01423X/1. The continuing support of Pembroke College, Oxford, is gratefully acknowledged.

LITERATURE CITED

Ba S, Joseph VR. 2012. Composite Gaussian process models for emulating expensive functions. *Ann. Appl. Stat.* 6:1838–60

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

Bayarri MJ, Berger JO, Calder ES, Dalbey K, Lunagomez S, et al. 2009a. Using statistical and computer models to quantify volcanic hazards. *Technometrics* 51:402–13

Bayarri MJ, Berger JO, Calder ES, Patra AK, Pitman EB, et al. 2015. Probabilistic quantification of hazards: a methodology using small ensembles of physics-based simulations and statistical surrogates. *Int. J. Uncertain. Quantif.* 5:297–325

Bayarri MJ, Berger JO, García-Donato G, Liu F, Palomo J, et al. 2007a. Computer model validation with functional output. *Ann. Stat.* 35:1874–906

Bayarri MJ, Berger JO, Kennedy MC, Kottas A, Paulo R, et al. 2009b. Predicting vehicle crashworthiness: validation of computer models for functional and hierarchical data. *J. Am. Stat. Assoc.* 104:929–42

Bayarri MJ, Berger JO, Paulo R, Sacks J, Cafeo JA, et al. 2007b. A framework for validation of computer models. *Technometrics* 49:138–54

Berger JO. 2013. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer

Berliner LM. 2003. Physical-statistical modeling in geophysics. *J. Geophys. Res. Atmos.* 108:8776

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus A* 60:663–78
Conti S, O'Hagan A. 2010. Bayesian emulation of complex multi-output and dynamic computer models. *J. Stat. Plann. Inference* 140:640–51
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: Volume III*, ed. C Gatsonis, JS Hodges, RE Kass, R McCulloch, P Rossi, ND Singpurwalla, pp. 37–94.
Currin C, Mitchell T, Morris M, Ylvisaker D. 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Am. Stat. Assoc.* 86:953–63
Du H, Smith L. 2012. Parameter estimation through ignorance. *Phys. Rev. E* 86:016213
Easterling RG. 2001. *Measuring the predictive capability of computational models: principles and methods, issues and illustrations.* Tech. Rep. SAND2001-0243, Sandia Natl. Lab., Albuquerque, NM
Fricker TE, Oakley JE, Urban NM. 2013. Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics* 55:47–56
Goldstein M, Rougier JC. 2003. *Calibrated Bayesian forecasting using large computer simulators.* Tech. Rep., Stat. Probab. Group, Univ. Durham, UK. **http://www.maths.dur.ac.uk/stats/physpred/papers/CalibratedBayesian.ps**
Goldstein M, Rougier JC. 2004. *Probabilistic formulations for transferring inferences from mathematical models to physical systems*. Tech. Rep., Stat. Probab. Group, Univ. Durham, UK. **http://www.maths.dur.ac.uk/stats/physpred/papers/directSim.pdf**
Gramacy RB, Lee HK. 2009. Adaptive design and analysis of supercomputer experiments. *Technometrics* 51:130–45
Gu M, Berger JO, et al. 2016a. Parallel partial Gaussian process emulation for computer models with massive output. *Ann. Appl. Stat.* 10:1317–47
Gu M, Palomo J, Berger J. 2016b. `RobustGaSP`: robust Gaussian stochastic process emulation. *R package version 0.5*. **https://cran.r-project.org/web/packages/RobustGaSP/index.html**
Gu M, Wang X, Berger JO. 2018. Robust Gaussian stochastic process emulation. *Ann. Stat.* 46:3038–66
Hagedorn R, Smith L. 2009. Communicating the value of probabilistic forecasts with weather roulette. *Meteorol. Appl.* 16:1749–72
Higdon D, Gattiker J, Williams B, Rightley M. 2008. Computer model calibration using high-dimensional output. *J. Am. Stat. Assoc.* 103:570–83
Hirsch M, Smale S. 1974. *Differential Equations, Dynamical Systems, and Linear Algebra*. New York: Academic
Howson C. 2000. *Hume's Problem: Induction and the Justification of Belief*. Oxford, UK: Oxford Univ. Press
Insua DR, Ruggeri F. 2012. *Robust Bayesian Analysis*. New York: Springer
Judd K, Smith L. 2004. Indistinguishable states II: imperfect model scenario. *Physica D* 196:224–42
Kennedy MC, O'Hagan A. 2000. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* 87:1–13
Kennedy MC, O'Hagan A. 2001. Bayesian calibration of computer models (with discussion). *J. R. Stat. Soc. B* 63:425–64
Le Gratiet L, Cannamela C, Iooss B. 2014. A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes. *J. Uncertain. Quantif.* 2:336–63
Levi I. 1983. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Cambridge, MA: MIT press
Linkletter C, Bingham D, Hengartner N, Higdon D, Kenny QY. 2006. Variable selection for Gaussian process models in computer experiments. *Technometrics* 48:478–90
Liu F, Bayarri MJ, Berger J. 2009. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* 4:119–50
Lorenz E. 1985. The growth of errors in prediction. In *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, ed. M Ghil, R Benzi, G Parisi, pp. 243–65. Amsterdam: North-Holland

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF Ensemble Prediction System: methodology and validation. *Q. J. R. Meteorol. Soc.* 122:73–119
Morris MD, Mitchell TJ, Ylvisaker D. 1993. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics* 35:243–55
NRC (Natl. Res. Counc.). 1979. *Carbon Dioxide and Climate: A Scientific Assessment*. Tech. Rep., Natl. Acad. Sci., Washington, DC
Oakley J, O'Hagan A. 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Stat. Soc. B* 66:751–69
Oberkampf WL, Trucano T. 2000. *Validation methodology in computational fluid dynamics*. Paper presented at Fluids 2000, June 19–22, Denver, CO
Ogburn SE, Berger J, Calder ES, Lopes D, Patra A, et al. 2016. Pooling strength amongst limited datasets using hierarchical Bayesian analysis, with application to pyroclastic density current mobility metrics. *Stat. Volcanol.* 2:1–26
Parker WS. 2011. When climate models agree. *Philos. Sci.* 78(4):579–600
Parker WS. 2014. Simulation and understanding in the study of weather and climate. *Perspect. Sci.* 22:336–56
Parker WS, Risbey JS. 2015. False precision, surprise and improved uncertainty assessment. *Phil. Trans. R. Soc. A* 373:20140453
Paulo R, García-Donato G, Palomo J. 2012. Calibration of computer models with multivariate output. *Comput. Stat. Data Anal.* 56:3959–74
Pilch M, Trucano T, Moya JL, Froehlich G, Hodges A, Peercy D. 2001. *Guidelines for Sandia ASCI verification and validation plans—content and format: Version 2.0.* Tech. Rep. SAND 2001-3101, Sandia Natl. Lab., Albuquerque, NM
Plumlee M. 2017. Bayesian calibration of inexact computer models. *J. Am. Stat. Assoc.* 112:1274–85
Qian PZG, Wu H, Wu CFJ. 2008. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* 50:383–96
Roache PJ. 1998. *Verification and Validation in Computational Science and Engineering*. Albuquerque, NM: Hermosa
Rougier J. 2008. Efficient emulators for multivariate deterministic functions. *J. Comput. Gr. Stat.* 17:827–43
Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989. Design and analysis of computer experiments. *Stat. Sci.* 4:409–23
Santner TJ, Williams BJ, Notz WI. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer
Savitsky T, Vannucci M, Sha N. 2011. Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Stat. Sci.* 26:130–49
Schonlau M, Welch WJ. 2006. Screening the input variables to a computer model via analysis of variance and visualization. In *Screening*, ed. A Dean, S Lewis, pp. 308–27. New York: Springer
Smith LA, 1995. Accountability and error in ensemble forecasting. In *1995 ECMWF Seminar on Predictability*, Vol. 1, pp. 351–68. Reading, UK: ECMWF
Smith LA. 1997. The maintenance of uncertainty. In *Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models*, ed. GC Castagnoli, A Provenzale, pp. 177–246. Bologna, Italy: Societá Italiana di Fisica
Smith LA. 2002. What might we learn from climate forecasts? *PNAS* 99:2487–92
Smith LA. 2016. Integrating information, misinformation and desire: improved weather-risk management for the energy sector. In *UK Success Stories in Industrial Mathematics*, ed. PJ Aston, AJ Mulholland, KMM Tant, pp. 289–96. New York: Springer
Smith LA, Petersen AC. 2015. Variations on reliability: connecting climate predictions to climate policy. In *Error and Uncertainty in Scientific Practice*, ed. M Boumans, G Hon, AC Petersen, pp. 151–70. London: Routledge
Smith LA, Stern N. 2011. Uncertainty in science and its role in climate policy. *Phil. Trans. R. Soc. A* 369:4818–41
Smith RC. 2014. *Uncertainty Quantification: Theory, Implementation, and Applications*. Philadelphia: SIAM
Smith RL, Tebaldi C, Nychka D, Mearns LO. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *J. Am. Stat. Assoc.* 104:97–116

Sobol IM. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* 55:271–80

Spiller ET, Bayarri M, Berger JO, Calder ES, Patra AK, et al. 2014. Automating emulator construction for geophysical hazard maps. *J. Uncertain. Quantif.* 2:126–52

Stainforth DA, Aina T, Christensen C, Collins M, Faull N, et al. 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* 433:403–6

Stocker T, Qin D, Plattner G, Tignor M, Allen S, et al., eds. 2013. *Climate Change 2013: The Physical Science Basis.* Cambridge, UK: Cambridge Univ. Press

Storlie CB, Lane WA, Ryan EM, Gattiker JR, Higdon DM. 2015. Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline ANOVA. *J. Am. Stat. Assoc.* 110:68–82

Tarantola S, Gatelli D, Kucherenko S, Mauntz W, et al. 2007. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Eng. Syst. Saf.* 92:957–60

Tennekes H, Baede APM, Opsteegh JD. 1986. *Forecasting forecast skill*. Presented at Workshop on Predictability in the Medium and Extended Range, March 17–19, Shinfield Park, Reading, UK

Trucano T, Pilch M, Oberkampf WO. 2002. *General concepts for experimental validation of ASCI code applications*. Tech. Rep. SAND 2002-0341, Sandia Natl. Lab., Albuquerque, NM

van den Dool H. 2007. *Empirical Methods in Short-Term Climate Prediction*. Oxford, UK: Oxford Univ. Press

Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD. 1992. Screening, predicting, and computer experiments. *Technometrics* 34:15–25