Studies in History and Philosophy of Modern Physics 41 (2010) 263–272

Contents lists available at ScienceDirect

Studies in History and Philosophy of Modern Physics

journal homepage: www.elsevier.com/locate/shpsb

Predicting weather and climate: Uncertainty, ensembles and probability

Wendy S. Parker [1]

Ohio University, Department of Philosophy, Ellis Hall 202, Athens, OH 45701, USA

ARTICLE INFO

Keywords:
Simulation
Climate
Weather
Uncertainty
Ensemble
Prediction

ABSTRACT

Simulation-based weather and climate prediction now involves the use of methods that reflect a deep concern with uncertainty. These methods, known as ensemble prediction methods, produce multiple simulations for predictive periods of interest, using different initial conditions, parameter values and/or model structures. This paper provides a non-technical overview of current ensemble methods and considers how the results of studies employing these methods should be interpreted, paying special attention to probabilistic interpretations. A key conclusion is that, while complicated inductive arguments might be given for the trustworthiness of probabilistic weather forecasts obtained from ensemble studies, analogous arguments are out of reach in the case of long-term climate prediction. In light of this, the paper considers how predictive uncertainty should be conveyed to decision makers.

1. Introduction

Computer simulation models are ubiquitous in the study of weather and climate. Their value as aids to understanding is widely acknowledged, but for prediction they are considered truly indispensable. Over the last half-century, substantial resources have been devoted to the development of these weather and climate models and to the expansion of observing networks that help to provide initial conditions for them, often with the explicit goal of more accurate prediction. Some of the fruit of this investment can be seen in the considerable increase in the skill of weather forecasts that has occurred in recent decades (e.g. Simmons & Hollingsworth, 2002).

It is important to recognize, however, that simulation-based weather and climate prediction today differs from that undertaken 50 years ago in ways that extend far beyond having more comprehensive observing systems and higher-resolution models that represent more physical processes in greater detail. Most notably, weather and climate prediction today involves the use of methods that reflect a deep concern with uncertainty. These methods, known as *ensemble prediction methods*, produce multiple simulations for predictive periods of interest, using different initial conditions, parameter values and/or model structures.

Ensemble methods are judged to be among the best ways forward when it comes to predicting weather and climate in the face of uncertainty. Yet interpreting results from studies employing ensemble methods is a complex matter and, in the context of climate prediction, is currently the subject of much discussion and debate (e.g. Collins, 2007; Stainforth, Allen, Tredger, & Smith, 2007). Particularly contentious is whether the results of ensemble prediction studies should be transformed into probabilistic forecasts that are offered as such to decision makers—as they choose policies for reducing greenhouse gas emissions, decide how high to build sea walls to protect coastal cities from strong storms, etc.

The discussion that follows has two main goals. First, it aims to provide a non-technical introduction to today's ensemble prediction methods in a way that gives a sense of the richness and complexity of current practice. As we will see, weather and climate prediction involves much more than simply "running a model". Second, the discussion will consider how results produced using the latest ensemble methods should be interpreted. A key conclusion will be that, while complicated inductive arguments might be given for the trustworthiness of probabilistic weather forecasts obtained from ensemble studies, analogous arguments are out of reach in the case of long-term climate prediction. In light of this, options for conveying predictive uncertainty to decision makers are presented.

2. Representational uncertainty in global weather and climate prediction

A computer simulation model is a computer-implemented set of instructions for repeatedly solving a set of equations in order to

E-mail address: parkerw@ohio.edu

[1] Thanks to Lenny Smith for helpful discussion, to Gavin Schmidt for providing Fig. 1, and to Tim Palmer and Royal Society Publishing for permission to use Fig. 2. This material is based on work supported by the National Science Foundation under Grant no. 0824287.

1355-2198/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsb.2010.07.006

produce a representation of the temporal evolution (if any) of selected properties of a target system. For simulation models used to forecast the weather, the target system is Earth's atmosphere, and the equations of interest are ones that specify the instantaneous rate of change of temperature, pressure, winds and humidity at any given location in the atmosphere. To arrive at these differential equations, the atmosphere is treated as a moving fluid in which various physical processes, such as the absorption and emission of radiation and the formation of clouds and precipitation, are also adding (or removing) energy and moisture locally. Since these equations cannot be solved analytically—they include, among other things, a version of the Navier–Stokes equations—solutions are estimated using numerical methods, including finite differencing techniques, with the help of a computer. Solutions might be produced in ~10 min time steps, for each of many points on a global spatial grid, with grid points spaced the equivalent of ~50–100 km in the horizontal on each of ~30 vertical levels (UCAR, 2009a).[2] In this way, a series of predictive snapshots of the global atmosphere is produced, typically extending about two weeks into the future.

For global climate models, the target system is Earth's climate system, understood to include the atmosphere, oceans, sea ice and land surface. When predicting future climate, the aim is not to simulate conditions in the climate system on any particular day, but to simulate conditions over a longer period—typically years or more—in such a way that the statistics of the simulated conditions (i.e. the simulated climate) will match the statistics of conditions that would occur during that period (i.e. the actual climate) under some specified emission scenario.[3] The atmosphere component of a global climate model is typically very similar to a weather forecasting model. Representations of the ocean, sea ice and land surface vary more from model to model, with some models including more simplified representations than others (e.g. a "slab" ocean as opposed to a fully dynamic ocean). Once again, some of the modeling equations are analytically intractable, and the computer is used to estimate solutions. Since climate simulations often cover decades or centuries of system evolution, the resolution of global climate models must be coarser than weather forecasting models if they are to finish running in time to be useful; solutions might be produced in ~30 min time steps in the atmosphere component of a complex climate model, with grid points spaced the equivalent of ~200 km in the horizontal on each of ~20–40 vertical levels.

Identifying appropriate initial conditions from which to run today's weather and climate models is not an easy task. A value must be assigned to each of the model's variables for the start time of the simulation period, t_i. For weather forecasting models, this will typically include values for temperature, pressure, wind speed and humidity for each of the model's many grid points. Climate models will often require, in addition, values for temperature, velocity, salinity and other variables for each of many ocean grid points, as well as values for all variables included in the sea ice and land surface components. Depending on the resolution of the models, i.e. on the mesh of the spatial grid on which solutions to model equations will be estimated, this can amount to 10^6 or 10^7 variables for which initial values are needed!

Observations of actual conditions in the atmosphere or ocean around time t_i are the natural starting point when choosing initial conditions. However, observations are subject to various errors, and they are made at locations that are irregularly and/or widely spaced. An activity known as *data assimilation* aims to remedy this. In general terms, data assimilation is a process by which information from multiple sources is combined in an attempt to estimate the state of a system for a time of interest (see Talagrand, 1997). For the atmosphere and ocean, the assimilation procedure often uses a model-produced forecast for time t_i as a first guess regarding actual conditions at t_i and then updates those forecasted values in light of observations made during a period of time extending on either side of t_i, known as the *assimilation window*.[4] Updating may factor in that both observations and model forecasts are subject to error. In this way, it is possible to produce a best-guess estimate of the state of the atmosphere (or oceans) for time t_i that includes values for all model variables at all atmosphere (or ocean) grid points.

In the case of weather forecasting, a best-guess estimate of the state of the atmosphere produced via data assimilation is referred to as an *analysis* and provides initial conditions with which to generate a forecast. For simulations of future climate, initialization methods are more varied. For some global climate models, the ocean component will be initialized with values obtained via data assimilation, while the atmosphere component is initialized with values obtained from a short atmosphere-only simulation; for other models, initial conditions for both components will be obtained from much longer simulations produced with the ocean and atmosphere components individually (Randall et al., 2007, p. 607). Initial conditions for all variables in the sea ice and land surface components of course also need to be specified using available information.

However initial conditions are chosen, there will be uncertainty associated with the choice; the state of the atmosphere or climate system might have been represented just as plausibly in a slightly different way. This looseness stems from the fact that both available observations and today's models are imperfect in ways that are neither readily apparent nor readily correctible. Uncertainty regarding the choice of initial conditions became a source of concern in the context of weather forecasting several decades ago, when Ed Lorenz famously discovered that even small differences in the conditions used to initialize weather models can lead to quite large differences in the forecasts produced (see Lorenz, 1963, 1965). Indeed, it was the recognition of this sensitive dependence on initial conditions that first prompted atmospheric scientists to consider ensemble approaches (Leith, 1974; Lorenz, 1965). The extent to which initial condition uncertainty is problematic in the context of climate prediction, where statistical features of simulations are of interest, remains unclear, but the existence of initial condition uncertainty is readily acknowledged.

There is also uncertainty associated with the choice of modeling equations, for several reasons. First, while the aim is to build weather and climate models from well-established physical theories, not all relevant physical processes are well understood from a theoretical point of view. Second, even processes that are well understood may need to be represented in a simplified or idealized way, because their theoretical equations are analytically unsolvable for the cases of interest or because estimating solutions numerically would be too complicated or too computationally demanding. Several simplified or idealized representations of a process may be developed, without it being obvious that one representation is better (for the

[2] This is for models using finite differencing methods of solution, rather than spectral methods. Note also that the resolution specified here is for global models; regional models typically have higher resolution.

[3] An *emission scenario* is an account of how greenhouse gas emissions and other human-related activities that can influence climate might evolve in the future.

[4] The assimilation window might be very short, extending just a few minutes on either side of t_i, or somewhat longer, extending hours or more, depending on the purpose for which the estimated state is to be used.

purposes at hand) than the others. Third, some important processes—such as the formation of individual clouds and the transfer of momentum by turbulence—occur on scales smaller than those resolved by global weather and climate models and thus cannot be simulated explicitly; they must be represented (somehow) in terms of larger-scale conditions. This activity is known as *parameterization*. In general, there is no obviously best way to parameterize a given sub-grid process.[5]

For all of these reasons, there is often considerable uncertainty about which modeling equations would be best, or even adequate, for predicting particular features of weather and climate with desired accuracy. Even with the substantial store of knowledge that scientists have accumulated regarding the atmosphere and climate system—knowledge that does lead to confident representations of some processes—at numerous points in the model-building process there is no obviously best way to do things. When speaking of this uncertainty about how to adequately represent the processes that shape the evolution of weather and climate, scientists today usually make a further distinction, between structural and parametric uncertainty. *Structural uncertainty* often refers to uncertainty about the form that modeling equations should take (e.g. should this quantity be represented as a function of just variable x, or of both variable x and variable y), while *parametric uncertainty* is uncertainty about the values that should be assigned to parameters within a set of modeling equations (e.g. in this equation, should this parameter be set to 0.2 or 0.3 m/s?).

Thus, there are three major types of representational uncertainty recognized in the context of global weather and climate simulation today: Initial condition uncertainty, structural uncertainty and parametric uncertainty.[6]

3. Building ensembles for weather and climate prediction

Ensemble methods take a brute-force approach to exploring the implications of the representational uncertainty just discussed. The basic idea is simple enough: Rather than run just one predictive simulation, run multiple simulations, sampling different initial conditions, parameter values or modeling equations that are plausibly adequate for the predictive task at hand. A collection of predictions will be produced, rather than just one, reflecting the representational uncertainty sampled by the alternative initial conditions, parameter values or modeling equations that are used.

This characterization of ensemble methods calls to mind Monte Carlo estimation, whereby probability distributions that reflect uncertainty regarding initial condition or parameter values are randomly sampled many times, and a simulation is produced with each set of values selected via the sampling, in order to estimate (in the form of another probability distribution) uncertainty in one or more output variables. And indeed, it seems fair to say that ensemble prediction of weather and climate is inspired by Monte Carlo methods (see e.g. Leith, 1974).

But ensemble studies of weather and climate differ from the paradigmatic Monte Carlo approach just described, for several reasons. One is the high dimensionality of the representational uncertainty at issue. Consider initial condition uncertainty: If there are $\sim 10^6$ variables for which values are needed, and some uncertainty about which value should be assigned to each, then to directly explore this uncertainty space a tremendous number of samples (and corresponding simulations) would be called for, vastly outstripping available computing power. Indeed, even one simulation of 21st century conditions using a state-of-the-art climate model can require significant time on a supercomputer, rendering computationally infeasible the running of many simulations. In addition, when it comes to structural uncertainty, it is not clear how to sample adequately from a space of mathematical functions in the way that one might from a space of numerical values, as in a traditional Monte Carlo study. Indeed, even characterizing a space of functions from which to sample can be quite difficult (see Murphy et al., 2007; Parker, in press).

The remainder of this section gives a non-technical overview of some of the ways in which scientists are producing ensembles of predictive simulations in the face of these conditions (high dimensionality, computational intensity and structural uncertainty), leaving discussion of the interpretation of ensemble results for Section 4.

3.1. Building ensembles for weather prediction

In December 1992, the National Centers for Environmental Prediction (NCEP) in the United States and the European Center for Medium-Range Weather Forecasting (ECMWF) in the United Kingdom became the first forecasting centers to implement ensemble methods as a regular part of daily weather prediction (Kalnay, 2003).[7] Since then, various other weather forecasting centers around the world, including the Meteorological Service of Canada (MSC), have also implemented ensemble methods.[8] The forecasting centers differ in the extent to which they focus on initial condition uncertainty as opposed to parametric and structural uncertainty and in how they investigate these different kinds of uncertainty.[9]

At NCEP and ECMWF, the focus has been on initial condition uncertainty. The forecast procedure begins with an analysis obtained via data assimilation. The analysis provides one set of initial conditions from which to produce a forecast, often referred to as the *control forecast*. Plausible alternative sets of initial conditions are generated by making small changes (or *perturbations*) to the analysis. At both centers, the aim has been to identify fast-growing perturbations—ones that will lead to particularly large differences in the forecasts produced. Nevertheless, the two centers take different approaches to identifying such fast-growing perturbations.

The *ensemble transform bred vector* approach used at NCEP (see Toth & Kalnay, 1997; Wei, Toth, Wobus, & Zhu, 2008) is backward-looking: It generates perturbations that emphasize the respects in which recently made forecasts for t_i differ most from one another. By contrast, the *singular value decomposition* approach employed at ECMWF (see Buizza et al., 2005) seeks perturbations that are expected on mathematical grounds to grow fastest in the near future.[10] Each approach has some advantages. The bred vector approach is low cost, requiring very little

[5] For a detailed look at the many equations included in the atmosphere component of one complex climate model, see UCAR (2009b).

[6] Uncertainty regarding the choice of boundary conditions in weather forecasting models—values for soil moisture, vegetation cover and other properties of the underlying land or ocean surface—is also relevant and is addressed by some forecasting groups (see e.g. Buizza et al., 2005). But initial condition uncertainty is given much more attention.

[7] NCEP was then known as the National Meteorological Center (NMC).

[8] This paper focuses on medium-range ensemble forecasting with global models. Ensemble approaches to short-range forecasting for sub-global regions with mesoscale models (e.g. Gel, Raftery, & Gneiting, 2004; Grimit & Mass, 2002) are also undergoing rapid development but are not discussed here, due to limitations of space.

[9] The discussion here provides a snapshot of some major recent approaches to ensemble prediction. Ensemble systems change frequently as computing power increases and new techniques are developed; this paper's characterizations of ensemble prediction systems at particular forecasting centers will likely be out of date in some respects by the time the paper appears.

[10] Ed Lorenz suggested an approach like this already in the 1960s (see Lorenz, 1965, pp. 331–332).

Table 1
Summary of means of accounting for initial condition, parametric and structural uncertainty at three major operational weather forecasting centers.

	Initial condition uncertainty	Parametric uncertainty	Structural uncertainty
ECMWF	Alternative initial conditions via singular value decomposition	Stochastic physics	Stochastic physics
MSC	Alternative initial conditions via ensemble transform Kalman filter	Alternative parameter values for turbulent vertical diffusion, gravity wave drag	Alternative parameterizations for convection, land surface processes, mixing length; stochastic physics
NCEP	Alternative initial conditions via ensemble transform bred vectors	Inflated initial condition uncertainty[a]	Inflated initial condition uncertainty[a]

[a] Approach taken as of July 2002 (see Buizza et al., 2005); no accounting of parametric and structural uncertainty is mentioned for NCEP in UCAR (2009a).

computation to produce alternative sets of initial conditions, because it draws upon forecasts already made. The singular value decomposition approach is more expensive computationally, but its results are determined by properties of the analysis for t_i, rather than by model behavior in the recent past; this would seem to be an advantage, since the perturbations that will grow fastest vary to some extent from day to day—they are state-dependent.

NCEP and ECMWF also give some attention to structural and parametric uncertainty, but they use indirect means. For instance, NCEP has allowed alternative sets of initial conditions to reflect perturbations that are somewhat larger than the estimated uncertainties in the analysis (Buizza et al., 2005, p. 1080), in effect inflating initial condition uncertainty in an attempt to produce the wider range of results that (presumably) would be obtained if alternative model structures and parameter values were explored directly. At ECMWF, structural and parametric uncertainties are addressed indirectly using *stochastic physics*, which involves perturbing physical tendencies at each time step during the simulation. For example, suppose it is calculated that parameterized sub-grid processes together would warm a locale by 0.03 °C during a given time step in the simulation. This calculated contribution from parameterized processes is multiplied by a number selected randomly from the interval [0.5, 1.5], and the result is used in place of the 0.03 °C originally calculated (see Buizza, Miller, & Palmer, 1999). Again, the aim is to produce the wider range of results that (presumably) would be obtained if alternative model structures and parameter values were explored directly.

With the help of supercomputers, these methods are used to produce 21 16-day global weather forecasts every 6 h at NCEP and to produce 51 15-day global weather forecasts every 12 h at ECMWF (ECMWF, 2006; UCAR, 2009a).

A rather different approach is taken at MSC in Canada, where 21 16-day global forecasts are produced every 12 h using not just alternative sets of initial conditions, but alternative forecast models as well. Sets of initial conditions are produced using an *ensemble transform Kalman filter* method, which performs data assimilation on perturbed observations using different versions of the MSC forecast model, producing 96 sets of initial conditions (UCAR, 2009a). The average of these sets of initial conditions is used with the MSC forecasting model in its standard configuration to produce a control forecast, while each of 20 sets of initial conditions (selected from the 96 sets produced) is paired with one alternative version of the MSC forecast model to produce 20 more forecasts (UCAR, 2009a). These alternative model versions—also used in the data assimilation process—differ in some of their parameter values and in how they parameterize some physical processes. For example, there is variation in the values assigned to parameters associated with gravity wave drag and turbulent vertical diffusion and also in the types of parameterizations used for convection and land surface processes (Environment Canada, 2007). The different parameter values and parameterizations are judged to be plausible alternatives to those used in the standard version of the model (Houtekamer and Lefaivre, 1997). So in contrast to the indirect approaches taken at NCEP and ECMWF, the ensemble prediction system implemented at MSC investigates parametric and structural uncertainty directly, by generating forecasts using different model versions. In addition, stochastic physics is employed, i.e. physical tendencies associated with parameterized processes are perturbed at each time step in the simulation (Table 1).

3.2. Building ensembles for climate prediction

Unlike short-term weather conditions, the climate of the next decade, or the next century, is not routinely predicted.[11] In part, this is because some of the factors that are believed to influence long-term climate—such as concentrations of greenhouse gases—depend on human activities; how these factors evolve over the next century, and thus how climate will change over the next century, depends to some extent on human decisions in that period. So climate prediction studies tend to be ones that investigate not how climate will actually change, but how climate would change under emission scenarios that are of scientific or societal interest.[12] If we could see how climate would change under different emission scenarios, we might conclude that some scenarios are to be avoided, while others are worth pursuing.

Climate models are used to investigate what climate would be like under particular emission scenarios. But as noted above, there is uncertainty about how to best (or even adequately) represent the climate system when undertaking such predictive tasks. So far, ensemble studies carried out to investigate this uncertainty usually fall into one of two categories: *Multi-model ensemble* studies and *perturbed-physics ensemble* studies. The former typically include models that differ in a host of ways—in some of the equations they use to represent climate system processes, in their numerical solution methods, in their spatiotemporal resolution, etc. Multi-model ensemble studies therefore probe structural uncertainty (and perhaps parametric uncertainty as well, depending on the particular models included in the ensemble). Perturbed-physics studies produce simulations using a single climate model, but assigning different values to uncertain parameters, in order to investigate the impacts of parametric uncertainty. In both multi-model and perturbed-physics studies, it is not uncommon for a few alternative sets of initial conditions to be used as well, but structural and parametric uncertainty, respectively, seem to be the primary concern.

[11] Seasonal prediction is not discussed here, due to space limitations.

[12] Because they investigate what would happen if a scenario were to be realized, predictive simulations from climate models are often referred to as *projections* of future climate.

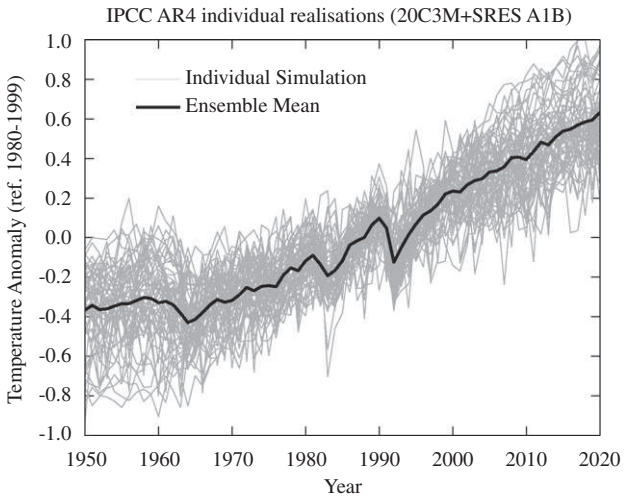

Fig. 1. Annual global mean surface temperature (GMST) anomalies from 55 individual CMIP3 simulations (grey) and the average of those anomalies (black). Anomalies for a given simulation are calculated relative to that simulation's average GMST for the period 1980–1999. Projections are for the A1B ("medium") emission scenario. Figure courtesy of Gavin Schmidt.

Perhaps the most ambitious multi-model ensemble study completed to date is the third phase of the Coupled Model Intercomparison Project (CMIP3), which produced simulations of future climate in support of the latest Intergovernmental Panel on Climate Change (IPCC) assessment report (Solomon et al., 2007). Twenty-three complex climate models developed at modeling centers around the world were used to simulate 21st century climate under three emission scenarios that represented futures with "high", "medium" and "low" global emission levels, respectively (see Meehl et al., 2007).[13] In some but not all cases, a model was run a few times under a given emission scenario, using different sets of initial conditions (Meehl et al., 2007, Table 10.4; see also Fig. 1 above). The CMIP3 models differed to varying degrees in their spatiotemporal resolution, their parameterization of sub-grid processes, their numerical solution techniques and their computing platforms (see Randall et al., 2007, Table 8.1). These differences were not systematically chosen, however; they were determined by which modeling groups agreed to participate in the study, by which models those groups had developed, and by the available computing resources. The CMIP3 models constitute an "ensemble of opportunity" (Meehl et al., 2007, p. 754; Tebaldi and Knutti, 2007).

The ongoing climateprediction.net project (Allen, 1999; Frame et al., 2009; Stainforth et al., 2005) provides fascinating examples of perturbed-physics studies. Rather than using supercomputers to produce a small number of higher-resolution simulations, the project relies on spare processing power on ordinary home computers to carry out ensemble studies that produce thousands of simulations using different versions of a somewhat lower-resolution (yet still quite complex) climate model. Any interested member of the public can participate in one of these studies by downloading a model version (+initial conditions) from the climateprediction.net website and running it on her home computer, with the results sent back automatically over the Internet. A single simulation may take several weeks to a few months to complete, depending on the computer used and on how much of its spare processing power is devoted to the simulation.

Climateprediction.net scientists are currently analyzing results from one of these perturbed-physics studies, carried out in cooperation with the British Broadcasting Corporation (BBC). The study, known as the BBC Climate Change Experiment, is investigating 21st century climate under the A1B/medium emission scenario that was also studied in CMIP3 (Frame et al., 2009). Thousands of versions of HadCM3L, a complex climate model developed at the UK Hadley Center, were downloaded by participants (BBC, 2009). Each model version included a unique combination of values for approximately 70 uncertain parameters, including parameters associated with the fall speed of ice crystals in clouds, the exchange of momentum between the ocean surface and the atmosphere, and the transfer of moisture from the soil to the atmosphere via plant transpiration; these combinations of parameter values all met a chosen plausibility requirement (Frame et al., 2009). In order to take some account of initial condition uncertainty, some participants received the same model version but one of several different sets of initial conditions.[14] In the end, the study produced tens of thousands of simulations of future climate under the chosen emission scenario.

Perturbed-physics studies also have been carried out with much more simplified climate models (e.g. Forest, Stone, Sokolov, Allen, & Webster, 2002; Knutti, Stocker, Joos, & Plattner, 2002; Meinshausen et al., 2009). These models require minimal computing time to run, allowing for studies that more closely approximate the ideal Monte Carlo approach described above. However, such models only simulate the evolution of highly aggregate quantities, such as hemispheric (or even global) mean surface temperature, and thus cannot be used to investigate the changes in regional climate that are of greatest interest to decision makers. In addition, because they are so simplified, there is concern that they may be incapable of predicting rapid changes in global climate produced by nonlinear feedbacks, which might be predicted by more complex climate models that represent such feedbacks explicitly.[15]

4. Interpreting ensemble results

Ensemble methods transform representational uncertainty into predictive uncertainty: A set of representations of the atmosphere or climate system is transformed into a set of predictive results. But what should be inferred from these results? This is a matter of some debate, especially in the case of climate prediction.

Least controversial is the following: Insofar as each model or model version (+initial conditions) in an ensemble is plausibly adequate for the predictive task of interest, then the simulations produced indicate a set of predictive outcomes that are plausible, given current knowledge.[16,17] If these outcomes vary widely, this implies substantial uncertainty regarding the future. While such a

[13] Most models were used to produce simulations for each of the three emission scenarios; some models were used to produce simulations for only some of the scenarios (see Meehl et al., 2007, Table 10.4).

[14] In some cases duplicate set-ups were assigned to different participants (Frame et al., 2009), presumably as a means of checking the experimental design (Stainforth et al., 2005).

[15] Some simple climate models can be tuned to emulate the known behavior of more complex models (behavior in simulating highly aggregate quantities, such as global mean surface temperature), raising interesting questions about the extent to which parametric and structural uncertainty can ultimately be distinguished (see also Meinshausen, Raper, & Wigley, 2008).

[16] This assumes no significant problems with the solution algorithm or computing platform.

[17] In the case of climate projections, these typically will be outcomes that are plausible under a chosen emission scenario, not necessarily plausible as outcomes that will actually occur.

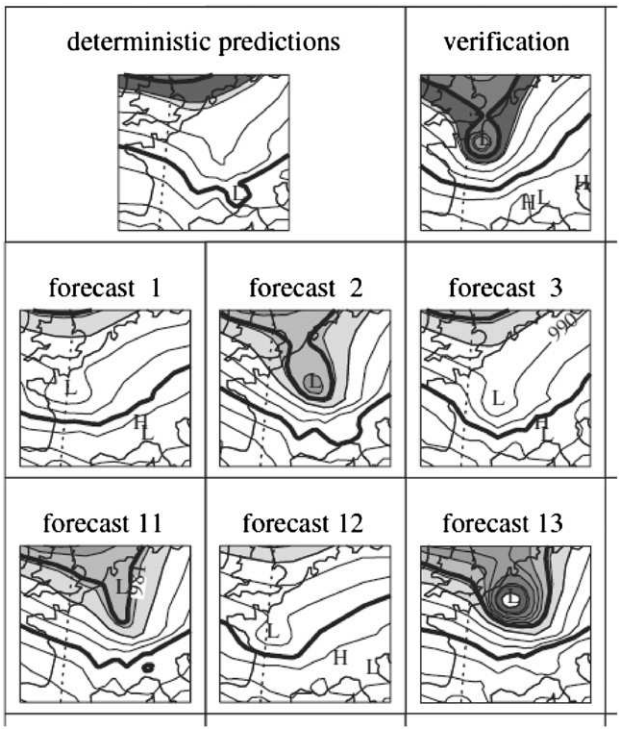

Fig. 2. An intense storm (top right, "verification") over Europe is predicted by some members of the ECMWF ensemble but not by the best-guess forecast (top left). Forecasts were for surface pressure on 26 December 1999, initialized on 24 December 1999. The destructive storm was later named "Lothar". Only seven of the 51 ECMWF ensemble members are shown here. (Adapted from Palmer, Doblas-Reyes, Hagedorn, & Weisheimer, 2005, Fig. 3.)

situation may be disappointing, it can be very useful to learn that, in addition to the outcome predicted by a best-guess/control simulation, a wide range of other outcomes are similarly plausible. Indeed, one of the great virtues of ensemble studies is that they can alert us to the existence of worrisome outcomes that are plausible but that did not occur in our best-guess simulation, such as the formation of a destructive winter storm in our region in a few days time (see Fig. 2) or very significant changes to the long-term climate of our region.

Yet it is tempting to infer more from ensemble results. If all of the climate models in an ensemble give predictions that fall within a rather narrow range, can we infer that there is little uncertainty about the predictive outcome? Or if 80% of simulations produced by an ensemble weather forecasting system indicate that there will be measurable snow tomorrow in Chicago, should we assign a probability of approximately 0.8 to the occurrence of measurable snow? More generally, can ensemble results be transformed in a systematic way into probability density functions (PDFs) whose implied probabilities for different outcomes can be expected to have a desired level of reliability, skill or utility?[18]

[18] A common measure of forecast quality is the Brier Skill Score, which compares the mean square error of a set of probabilistic forecasts for an event with the mean square error of a set of reference forecasts, typically forecasts of the climatological mean frequency of the event (see Toth, Talagrand, Candille, & Zhu, 2003). It can be shown that the Brier Skill Score is higher to the extent that forecasts have reliability and resolution. Roughly, forecasts are *reliable* (or *calibrated*) if their assigned probabilities for an event are consistent with observed relative frequencies, and they are *resolving* (or *sharp*) to the extent that they assign probabilities that differ from the climatological mean frequency of the event (see Stephenson, 2003, pp. 210–212). Other measures of forecast quality advocated recently include the ignorance score (Roulston & Smith, 2002) and the effective daily interest rate (Hagedorn & Smith, 2008).

There are at least two possible approaches to arguing that an ensemble study can be expected to provide desired information, including probabilistic information. The first approach is deductive and focuses on the design of the study, including the selection of models and initial conditions, and argues that this design is such that the results can be expected to provide the desired information. The second approach is inductive and argues that an ensemble system has been successful enough at providing the desired sort of information on relevant past occasions that it can be expected to do so in the present case(s) as well. It is suggested in what follows that, while design-based arguments generally cannot be given for claims about the information that today's ensembles can be expected to provide, performance-based arguments might be made in the case of ensemble weather prediction, but remain out of reach in the case of ensemble climate prediction.

4.1. Weather: Post-processing and complicated inductive arguments

Are any of today's ensemble weather forecasting systems designed such that (i) for a wide range of predictive variables, we can expect that the true/observed value of the variable will almost always fall within the range of values predicted by the ensemble? Are any of today's ensemble weather forecasting systems designed such that (ii) for a wide range of possible events, we can expect the fraction of ensemble members that predict a given event to be a reliable estimate of the probability of that event (e.g. 7/21 simulations gives a probability ≈ 0.33)? Unfortunately, it seems not.

A design-based argument for (i) would require that an ensemble forecasting system sample so much of current uncertainty about how to represent the atmosphere, or else sample that uncertainty in such a strategic way, that we can expect its predictions to almost always span a range of values that includes the true/observed value. But this cannot be claimed of today's ensemble forecasting systems. Perhaps the most obvious problem is their treatment of parametric and structural uncertainty; methods currently used to account for parametric and structural uncertainty (see Section 3.1) cannot be argued to sample that uncertainty in a sufficiently thorough or strategic way. The most direct approach is taken at MSC, where the forecasting system includes model versions with a few differences in their parameter values and structures. But these differences were not chosen with the aim of spanning some space of plausible options, if such a space can even be defined.

The most straightforward design-based argument for (ii) is thwarted for similar reasons. It would require that simulations produced by an ensemble forecasting system are, by design, something like a random sample from the simulations that would be produced in a thorough investigation of representational uncertainty. Yet today's ensemble prediction systems clearly have not been designed in this way, since the space of models (from which the sampling would occur) has not even been defined. Moreover, as noted above, it is not entirely clear what such a space would amount to, given structural uncertainty.

Can performance-based arguments for (i) and (ii) be given instead? These would require not just that an ensemble forecasting system be found to almost always capture observations and to give reliable probabilistic forecasts in a set of past cases, but also that there is reason to expect that the forecasting system will perform at least as well in the future cases of interest as it did in those past cases.

Performance-based arguments for (i) and (ii) cannot be given for raw results from today's ensembles, since those results are typically found to display bias and underdispersion (see e.g.

Buizza et al., 2005; Wilks & Hamill, 2007).[19] Nevertheless, there is currently great interest in *post-processing* ensemble results, i.e. in using information about the past performance of an ensemble to try to correct for biases and spread deficiencies and, ideally, to transform raw ensemble results into PDFs that are highly informative (e.g. Bröcker & Smith, 2007; Wilks & Hamill, 2007). Some post-processing is already done at operational forecasting centers, often using relatively simple methods, e.g. applying a bias correction in light of recent errors, with probabilistic forecasts then produced according to the fraction of bias-corrected ensemble members that predict the outcome of interest (UCAR, 2009a; see Environment Canada, 2009 for current forecasts). In some cases, even simple bias-correction techniques can substantially improve the performance of an ensemble prediction system. Recently scientists have demonstrated that a number of more complex post-processing methods, some of which deliver full PDFs, have the potential to further improve probabilistic forecasts for temperature, precipitation and other quantities (see e.g. Bröcker & Smith, 2007; Hagedorn, Hamill, & Whitaker, 2008; Hamill, Whitaker, & Mullen, 2006; Wilson, Beauregard, Raftery, & Verret, 2007).

So even if performance-based arguments for (i) and (ii) cannot be given for raw results from today's ensembles, they might be given for results that have undergone post-processing, if that post-processing is effective enough. To the extent that (i) and (ii) still prove too demanding, it might be possible to argue for weaker versions, such as (i′) when the value of variable X is predicted at lead time L, it can be expected that the observed value of X will fall within the range of bias-corrected results produced by this ensemble on about $p\%$ of occasions, or (ii′) for probabilistic forecasts of events of type E made at lead time L, this ensemble prediction system can be expected to display approximately skill S. Again, performance-based arguments would require that an ensemble prediction system is found to achieve the levels of performance specified in (i′) and (ii′) in a set of past cases and that there is good reason to expect it to perform similarly (if not better) in the future cases of interest.

But can the latter requirement ever be met? That is, can we ever have good reason to expect that a forecasting system will continue to perform in (approximately) some specified way in the future? In the spirit of Norton's (2003) material theory of induction, it is suggested here that there can be cases in which we have good reason to expect this, given what we know about the particular forecasting system and the predictive variables of interest. For instance, if we know that a forecasting system already has a long and stable track-record of performance with respect to a predictive variable, and we have reason to believe that neither the forecasting system nor the relevant causes at work in the atmosphere are now different in ways that will significantly alter the forecasting system's performance with respect to that predictive variable, then it seems we do have a case for expecting roughly that same performance to continue.

To illustrate: Suppose that in each of the last eight summers a particular ensemble system has delivered daily forecasts of the probability of E: The temperature in San Diego will exceed 30 °C on at least one of the next three days. Suppose further that in each of those summers the forecasts of E had a Brier Skill Score between 0.44 and 0.48, with a mean of 0.46. If we have reason to believe that neither the forecasting system nor the causes of hot days in San Diego will be different this summer in ways that will significantly alter the forecasting system's performance (relative to the previous eight summers), then we have reason to expect that this summer the forecasts of E will have a Brier Skill Score of approximately 0.46, i.e. will be in the vicinity of 0.44–0.48. Our expectation is grounded not just in the stability of past performance—we are not making a simple enumerative induction—but also in the domain-specific background knowledge we employ in concluding that there is nothing special about this summer's forecasts, i.e. that the sorts of things that would make for anomalous forecast performance (relative to performance in the last eight summers) are absent.[20] The strength of our inductive inference depends on the strength of that background knowledge. To the extent that we can confidently identify what would make for anomalous forecast performance (relative to performance in the last eight summers) and show those factors to be absent, our inference is stronger; to the extent that we have little idea what would make for anomalous forecast performance or whether such factors are present, our inference is weaker.

In practice, when it comes to today's ensembles, some additional complications arise. First, because ensemble forecasting systems (including the observing systems that provide input to data assimilation) undergo frequent development and change, most available performance data is not for today's systems but for earlier versions of those systems. This problem might be partly overcome with the help of "reforecasts"—forecasts produced for past periods using current models and current methods for generating alternative initial conditions (see e.g. Hamill et al., 2006). Alternatively, scientists might freeze today's forecasting systems and continue to make forecasts with them, even as they also develop and use new versions, so that eventually there is a significant track-record of performance for today's versions.[21] Second, there is the complication that Earth's climate is thought to be slowly changing over time due to rising greenhouse gas concentrations and other factors. In many cases, this might not undermine expectations of approximate stability of performance in the near term, but it should be considered, and it generally will be problematic if one wants to form expectations about performance in the further future.

These complications notwithstanding, in the end it seems quite plausible that performance-based arguments for today's ensembles could be developed for at least some instantiations of (i′) and (ii′) and perhaps also for other claims about the information that today's ensemble weather forecasting systems can provide. Whether such arguments can be given for aspects of performance that are relevant to important practical decisions (e.g. decisions about when to evacuate a region to avoid hurricane-related deaths) remains to be seen.

In this connection, notice that (i), (ii), (i′) and (ii′) all concern the expected performance of an ensemble forecasting system in a set of trials. It may seem natural to make assignments of probability in single cases as well.[22] For instance, if it can be expected that, over a given set of trials, an ensemble prediction system will give (approximately) reliable forecasts of the probability of event E, then following any one of those trials it may seem natural to assign to E a probability equal to that forecasted by the ensemble system. Thus, when the probability of snow

[19] A set of predictions for a variable is *biased* if its mean value tends to be greater than the observed value of the variable or tends to be less than that observed value. A set of predictions for a variable is *underdispersive* if observed values of the variable do not fall within the range spanned by the ensemble's results as often as would be expected if the results and the observations were being drawn from the same underlying PDF.

[20] In effect we are choosing a reference class for this set of forecasts and arguing on meteorological (and perhaps statistical) grounds that our choice is a good one, i.e. that there are no other features of this set of forecasts that are relevant to its classification for purposes of determining expected performance.

[21] This would not avoid changes in the observing systems that provide input to data assimilation.

[22] Some frequentists deny that it makes sense to assign probabilities to single cases.

tomorrow in Chicago is forecasted to be 0.8, a probability of about 0.8 would be assigned. But care must be taken here. The ensemble system gives approximately reliable forecasts of the probability of snow on the following day in Chicago; but if the event being forecasted is better classified in some other way, e.g. as the occurrence snow on the following day in Chicago when it is already snowing today in Des Moines, then a probabilistic forecast from an ensemble system that is expected to give reliable forecasts for this other class of events should be preferred instead, if it is available.[23] Further challenges will arise if two or more ensemble prediction systems are expected to give reliable forecasts for events in the chosen class, since they may nevertheless assign different probabilities in the particular case at hand.

4.2. Climate: Doing our best with uncertainty

When it comes to ensemble climate prediction today, neither design-based arguments for (i) or (ii) nor performance-based arguments for instantiations of (i′) or (ii′) can be given. Design-based arguments for (i) and (ii) are out of reach for the same reasons as in the weather case—today's ensembles are not designed to sample representational uncertainty in a thorough or strategic way. Today's multi-model ensembles are ensembles of opportunity, while perturbed-physics ensembles take no account of structural uncertainty. And in both multi-model and perturbed-physics studies, initial condition uncertainty is often given only cursory treatment.[24]

Performance-based arguments of the sort discussed above are out of reach in the first instance because of the long-term nature of the climate predictions of interest: It will take a very long time to collect much data on the performance of an ensemble in predicting the probability that a particular change in climate will have occurred after 50 years. And we do not have centuries and centuries of high-quality observations of past climatic conditions on which to evaluate performance either. (If such observational data were available, we might opt for reforecasting.) Simulations of 20th century climate can be compared with 20th century observational data, but the extent to which such simulations have been tuned to those data is often unclear, complicating attempts to reason about the expected performance of today's ensembles in simulating 21st century climate, when greenhouse gas concentrations are expected to be substantially higher.[25] Some estimates of climatic conditions in the more distant past are also available (from tree rings, ice cores, etc.), but there is more uncertainty associated with these data, and greenhouse gas concentrations in past periods are often lower than those anticipated for the 21st century. In the end, the available data—in conjunction with current understanding of the climate system—are not sufficient to support the kind of performance-based arguments discussed above.[26]

Nevertheless, it is becoming more and more common for results from individual multi-model and perturbed-physics studies to be transformed into probabilistic projections of future climate, using Bayesian and other techniques (e.g. Furrer, Sain, Nychka, & Meehl, 2007; Meinshausen et al., 2009; Murphy et al., 2007; Tebaldi, Smith, Nychka, & Mearns, 2005). In such cases, a set of simulation results is transformed into a PDF indicating which values of a given predictive variable are more and less probable. For instance, the PDF might imply that, under a chosen emission scenario, there is a probability of approximately 0.5 that global mean surface temperature in the period 2080–2099 will exceed that of the period 1980–1999 by at least 2 °C.

The reliability of these probabilistic projections is unknown, and in many cases they lack robustness. For instance, PDFs produced in different ensemble studies for the same quantity, e.g. for changes in global mean surface temperature by the end of the 21st century under a given emission scenario, can differ markedly (see Meehl et al., 2007, Fig. 10.28). Moreover, given the highly contingent way in which some of today's ensembles are assembled, the models that are included "might be different in a subsequent ensemble, therefore changing the result even if the knowledge about the climate system has not changed" (Tebaldi & Knutti, 2007, p. 2068). Thus, some scientists who carry out ensemble studies of future climate either refuse to transform raw results into PDFs via post-processing algorithms (see Stainforth et al., 2005, 2007) or else insist that, if PDFs are produced, they should be accompanied by an estimate of the chance of a "big surprise"—an outcome significantly outside the range into which the PDF implies the outcome is almost certain to fall (Smith, 2009).

It is important to recognize, however, that full PDFs are not the only option for conveying uncertainty about changes in future climate. For instance, Kandlikar, Risbey, and Dessai (2005) identify several ways of representing uncertainty: In the best case, when sources of uncertainty are well understood, it can be appropriate to convey uncertainty via full PDFs, but in other cases it will be more appropriate to offer only a range in which one expects the value of a predictive variable to fall with some specified probability, or to indicate the expected sign of a change without assigning a magnitude, etc. On their view, uncertainty should be expressed using the most "precise" means that can be justified, but not more precise means (ibid; see also Risbey & Kandlikar, 2007).

Building on this, it might be argued that, whatever depiction of uncertainty scientists offer to decision makers for a predictive variable of interest, it should meet three requirements: *Ownership, justification* and *robustness* (Parker, in press). In brief, this means that the representation of uncertainty should be one that the scientists will claim as their own, i.e. as accurately depicting their uncertainty (*ownership*); for which the scientists can offer a reasoned justification, after striving to consider all of the available, relevant evidence (*justification*); and that is not strongly dependent on contentious assumptions or expected to change significantly in the very near future as incremental scientific progress occurs (*robustness*). For most PDFs produced in today's ensemble climate prediction studies, especially those concerning long-term changes, at least one of these requirements remains unmet (see Parker, in press).

To see a concrete example of an alternative to a full PDF, consider Table 2, adapted from the latest IPCC assessment report (IPCC, 2007). For each of three emission scenarios, the table shows

[23] This is the reference class problem again (see Hajek, 2007). As in the case of a particular set of forecasts (see footnote 20), our task here is to argue on combined meteorological and statistical grounds that a particular classification of an event of interest is best. If classes A and B are both possibilities, and there is an ensemble system available that is expected to give reliable forecasts for events in class A, but there is currently no such system available for events in class B, then class A is the better choice. If for each class there is an ensemble system that is expected to give reliable forecasts for events in that class, and class B is a proper subset of class A, then class B is the better choice. In other situations, the choice may be more difficult.

[24] Ensemble studies with very simplified climate models might seem promising candidates for design-based arguments for (i) or (ii). However it is difficult to tell the extent to which these studies actually probe structural uncertainty. See also footnote 15.

[25] Tuning involves ad hoc adjustments to a model to achieve better fit with observational data or with the performance of another model.

[26] Of course, comparing simulations from today's ensembles with available climatic data can still be quite valuable, since we may learn about current strengths and weaknesses of the ensembles and may come to see how to improve our models. See also Smith, 2002.

Table 2
Estimated change in global mean surface temperature (late 21st century vs. late 20th century) under different emission scenarios. (Adapted from IPCC, 2007, Table SPM.3).

Temperature change (°C at 2090–2099 relative to 1980–1999)		
Scenario	**Best estimate**	***Likely* range**
B1 (low)	1.8	1.1–2.9
A1B (medium)	2.8	1.7–4.4
A2 (high)	3.4	2.0–5.4

a best estimate of the change in global mean surface temperature that would occur, as well as a *likely* range. The best estimate was chosen to correspond to the average of the values predicted by the state-of-the-art models included in CMIP3, while the *likely* range is that into which the experts producing the report judged there to be at least a 66% chance that the warming would fall (see Meehl et al., 2007 for details). Crucially, the experts identified these ranges by considering not just results from a single ensemble study, but results from numerous studies, factoring in as well that these studies had various known shortcomings (see Meehl et al., 2007, pp. 809–810). In the end, rather than offering a full PDF, the experts opted to report where some of the probability mass should be placed—namely, at least 66% of it.[27]

Few quantities are as well understood as changes in global mean surface temperature. For others, such as changes in regional precipitation, even *likely* ranges might be difficult to justify; in these cases, it would seem better to use a still coarser means of conveying uncertainty, e.g. reporting only an order of magnitude estimate of the change, or just its expected sign, or even that its expected sign is ambiguous (see also Kandlikar et al., 2005; Risbey & Kandlikar, 2007).[28] The aim should be to offer depictions of uncertainty that are as responsive as possible to the needs of decision makers but that also accurately reflect the limits of current knowledge.

5. Concluding remarks

Ensemble methods now play a central role in simulation-based weather and climate prediction. These methods acknowledge representational uncertainty and seek to gauge its predictive implications, going beyond a simple best-guess forecast or projection. However, interpreting the results of ensemble studies remains a challenging task, with probabilistic interpretations particularly contentious. While complicated inductive arguments might be made for the trustworthiness of some probabilistic weather forecasts produced in ensemble studies, the same cannot be said for PDFs produced for long-term climate variables. Consequently, in many cases, alternative means of communicating uncertainty about future changes in climate should be employed instead. The aim should be to offer depictions of uncertainty that are as responsive as possible to the needs of decision makers but that also accurately reflect the limits of current knowledge.

References

Allen, M. R. (1999). Do-it-yourself climate prediction. *Nature*, *401*, 642.

British Broadcasting Corporation (BBC). (2009). *Climate change experiment results*. Available from BBC website: ⟨http://www.bbc.co.uk/sn/climateexperiment/⟩; accessed 31.12.09.

Bröcker, J., & Smith, L. A. (2007). From ensemble forecasts to predictive distribution functions. *Tellus A*, *60*, 663–678.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., & Zhu, Y. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, *133*, 1076–1097.

Buizza, R., Miller, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*, 2887–2908.

Collins, M. (2007). Ensembles and probabilities: A new era in the prediction of climate change. *Philosophical Transactions of the Royal Society A*, *365*, 1957–1970.

Environment Canada. (2007). *Definition of the control model and perturbed models*. Available from Environment Canada website: ⟨http://www.weatheroffice.gc.ca/ensemble/verifs/model_e.html⟩; accessed 31.12.09.

Environment Canada. (2009). *Canadian ensemble forecasts*. Available from Environment Canada website: ⟨http://www.weatheroffice.gc.ca/ensemble/index_e.html⟩; accessed 31.12.09.

European Center for Medium-Range Weather Forecasting (ECMWF). (2006). *Implementation of VarEPS*. Available from ECMWF website: ⟨http://www.ecmwf.int/products/changes/vareps/⟩; accessed 30.12.09.

Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R., & Webster, M. D. (2002). Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, *295*, 113–117.

Frame, D. J., Aina, T., Christensen, C. M., Faull, N. E., Knight, S. H. E., & Piani, C. (2009). The climateprediction.net BBC climate change experiment: Design of the coupled model ensemble. *Philosophical Transactions of the Royal Society A*, *367*, 855–870.

Furrer, R., Sain, S. R., Nychka, D., & Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics*, *14*, 249–266.

Gel, Y., Raftery, A. E., & Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *Journal of the American Statistical Association*, *99*, 575–583.

Grimit, E. P., & Mass, C. F. (2002). Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather and Forecasting*, *17*, 192–205.

Hagedorn, R., & Smith, L. A. (2008). Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*. doi:10.1002/met.92.

Hagedorn, R., Hamill, T. M., & Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review*, *136*, 2608–2619.

Hajek, A. (2007). The reference class problem is your problem too. *Synthese*, *156*, 563–585.

Hamill, T. M., Whitaker, J. S., & Mullen, S. L. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, *87*, 33–46.

Houtekamer, P. L., & Lefaivre, L. (1997). Using ensemble forecasts for model validation. *Monthly Weather Review*, *125*, 2416–2426.

IPCC. (2007). *Summary for policymakers*. In Solomon et al. (Eds.) (pp. 1–18).

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. New York: Cambridge University Press.

Kandlikar, M., Risbey, J. S., & Dessai, S. (2005). Representing and communicating deep uncertainty in climate change assessments. *Comptes Rendus Geoscience*, *337*, 443–455.

Knutti, R., Stocker, T. F., Joos, F., & Plattner, G. K. (2002). Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, *416*, 719–723.

Leith, C. E. (1974). Theoretical skill of Monte-Carlo forecasts. *Monthly Weather Review*, *102*, 409–418.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*, 130–141.

Lorenz, E. N. (1965). Predictability of a 28-variable atmospheric model. *Tellus A*, *17*(3), 321–333.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., et al. (2007). Global climate projections. In Solomon et al. (Eds.) (pp. 747–845).

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., & Knutti, R. (2009). Greenhouse-gas emission targets for limiting global warming to 2 °C. *Nature*, *458*, 1158–1162.

Meinshausen, M., Raper, S. C. B., & Wigley, T. M. K. (2008). Emulating IPCC AR4 atmosphere–ocean and carbon cycle models for projecting global-mean, hemispheric and land/ocean temperatures: MAGICC 6.0. *Atmospheric Chemistry and Physics Discussions*, *8*, 6153–6272.

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

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*, 647–670.

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., & Weisheimer, A. (2005). Probabilistic prediction of climate using multi-model ensembles: From basics to applications. *Philosophical Transactions of the Royal Society B*. doi:10.1098/rstb.2005.1750.

[27] It is worth noting that, for a given scenario, many full PDFs produced in individual ensemble studies assigned at least 90% of the probability mass to a range smaller than that specified in Table 2 (see Meehl et al., 2007, Fig. 10.29).

[28] Further creativity here—in developing novel ways of depicting and communicating uncertainty—could be of significant value.

Parker, W. S. (in press). Whose probabilities? Predicting climate change with ensembles of models. In *PSA2008: Proceedings of the Philosophy of Science Association*.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., et al. (2007). *Climate models and their evolution*. In Solomon et al. (Eds.) (pp. 589–662).

Risbey, J. S., & Kandlikar, M. (2007). Expressions of likelihood and confidence in the IPCC assessment process. *Climatic Change*, *85*, 19–31.

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660.

Simmons, A. J., & Hollingsworth, A. (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *128*(580), 647–677.

Smith, L. A. (2002). What might we learn from climate forecasts?. *Proceedings of the National Academies of Science 99*, 2487–2492.

Smith, L. A. (2009). *Toward decision-relevant probability distributions: Communicating ignorance, uncertainty and model-noise* [*Powerpoint slides*]. Available from the Royal Meteorological Society website: 〈http://www.rmets.org/pdf/presentation/20091015-smith.pdf〉; accessed 13.01.10.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., & Miller, H. L. (Eds.). (2007). *Climate change 2007: The physical science basis. contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. New York: Cambridge University Press.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., & Frame, D. J. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*, 403–406.

Stainforth, D. A., Allen, M. R., Tredger, E. R., & Smith, L. A. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A*, *365*, 2145–2161.

Stephenson, D. (2003). Glossary. In I. Jolliffe, & D. Stephenson (Eds.), *Forecast verification: A practitioner's guide in atmospheric science* (pp. 203–213). Chichester: Wiley.

Talagrand, O. (1997). Assimilation of observations: An introduction. *Journal of the Meteorological Society of Japan*, *75*, 191–205.

Tebaldi, C., Smith, R., Nychka, D., & Mearns, L. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach. *Journal of Climate*, *18*, 1524–1540.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A*, *365*, 2053–2075.

Toth, Z., & Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, *125*, 3297–3319.

Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). Probability and ensemble forecasts. In I. Jolliffe, & D. Stephenson (Eds.), *Forecast verification: A practitioner's guide in atmospheric science* (pp. 137–163). Chichester: Wiley.

University Corporation for Atmospheric Research (UCAR). (2009a). *Operational models matrix: Characteristics of NWP and related forecast models*. Available from MedEd website: 〈http://www.meted.ucar.edu/nwp/pcu2/ens_matrix/index.htm〉; accessed 30.12.09.

University Corporation for Atmospheric Research (UCAR). (2009b). *Description of the NCAR community atmosphere model (CAM3)*. Available from University Corporation for Atmospheric Research website: 〈http://www.ccsm.ucar.edu/models/atm-cam/docs/description/〉; accessed 29.12.09.

Wei, M., Toth, Z., Wobus, R., & Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, *60*, 62–79.

Wilks, D. S., & Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, *135*, 2379–2390.

Wilson, L. J., Beauregard, S., Raftery, A. E., & Verret, R. (2007). Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, *135*, 1364–1385.