PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2007) **365**, 2145–2161
doi:10.1098/rsta.2007.2074
Published online 14 June 2007

Confidence, uncertainty and decision-support relevance in climate predictions

By D. A. Stainforth[1,3,*], M. R. Allen[2], E. R. Tredger[3] and L. A. Smith[3]

[1] *Tyndall Centre for Climate Change Research, Environmental Change Institute, Centre for the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK*
[2] *Department of Atmospheric, Oceanic and Planetary Physics, Oxford University, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK*
[3] *Centre for the Analysis of Time-series, Department of Statistics, Columbia House, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK*

Over the last 20 years, climate models have been developed to an impressive level of complexity. They are core tools in the study of the interactions of many climatic processes and justifiably provide an additional strand in the argument that anthropogenic climate change is a critical global problem. Over a similar period, there has been growing interest in the interpretation and probabilistic analysis of the output of computer models; particularly, models of natural systems. The results of these areas of research are being sought and utilized in the development of policy, in other academic disciplines, and more generally in societal decision making. Here, our focus is solely on complex climate models as predictive tools on decadal and longer time scales. We argue for a reassessment of the role of such models when used for this purpose and a reconsideration of strategies for model development and experimental design. Building on more generic work, we categorize sources of uncertainty as they relate to this specific problem and discuss experimental strategies available for their quantification. Complex climate models, as predictive tools for many variables and scales, cannot be meaningfully calibrated because they are simulating a never before experienced state of the system; the problem is one of extrapolation. It is therefore inappropriate to apply any of the currently available generic techniques which utilize observations to calibrate or weight models to produce forecast probabilities for the real world. To do so is misleading to the users of climate science in wider society. In this context, we discuss where we derive confidence in climate forecasts and present some concepts to aid discussion and communicate the state-of-the-art. Effective communication of the underlying assumptions and sources of forecast uncertainty is critical in the interaction between climate science, the impacts communities and society in general.

Keywords: climate change; uncertainty; probability; predictions; model inadequacy

[*] Author for correspondence (das@atm.ox.ac.uk).

One contribution of 13 to a Theme Issue ‘Ensembles and probabilities: a new era in the prediction of climate change’.

1. Introduction

The reality of anthropogenic climate change is well documented and widely accepted. The media and policy makers are calling out for predictions regarding expected changes to their local climate. Providing direct quantitative answers to these calls is perceived as important for engaging the public in the issue and therefore the task of mitigation. It is also often seen as critical for adaptation and decision making by businesses, governments and individuals. The extent to which these calls can be answered today is unclear, given the state of the science. Thus, the realistic communication of scientific uncertainty and the relevance of today's 'best available information' may prove critical for maintaining credibility in the future as model-based information improves.

A number of methods have been employed to provide detailed information about future climate. Sometimes, the output of one or more climate models, simulating one or more scenarios of future greenhouse gas (GHG) emissions, is simply taken as indicative and therefore suitable for purpose (e.g. UKCIP02 2002; Hayhoe *et al.* 2006). Methods have been proposed to create probability distributions of various sorts. Tebaldi *et al.* (2005) provide a Bayesian analysis of multiple models on regional scales as a representation of the belief (Giorgi & Mearns 2002) that greater inter-model similarity in simulated futures, combined with closer model/observation comparisons in simulated pasts, provides greater probability of an accurate forecast. Stott *et al.* (2006) apply a technique of scaling model predictions according to their ability to represent observed trends in the region in question. We note that Lopez *et al.* (2006) have demonstrated that these two methods give very different results even at the global scale. There is also the promise, by 2008, of probability distributions on up to 25 km spatial scales for multiple variables (UKCIP08 2006) based on processed output from perturbed physics ensembles (Allen & Stainforth 2002; Murphy *et al.* 2004).

How are we to determine the extent to which a given generation of climate models can provide decision-relevant information? At which spatial and temporal scales can the output of these models be interpreted as informative?

For changes in global mean temperature consistency with simpler methods such as those using energy balance models (e.g. Frame *et al.* 2005) supports the robustness of the distributions produced. Even these may be subject to assumptions about changes in ocean circulation and therefore effective ocean heat capacity. For virtually all other variables and spatial scales, these critical questions remain open. They require us to consider the uncertainties in the interpretation of climate models as predictions of the future.

There is no compulsion to hold that the most comprehensive models available will yield decision-relevant probabilities, even if those models are based upon 'fundamental physics'. Further, well-known fundamental barriers to the validation and verification of computer models of natural systems (Oreskes *et al.* 1994) remain a focus of ongoing research. There are many theoretical sources of uncertainty in the interpretation of model results in terms of the real world. Their separation and categorization is not always clear (Kennedy & O'Hagan 2001; O'Hagan & Oakley 2004). Here, we eschew a complete theoretical breakdown in favour of a discussion of those sources of uncertainty which we consider to be critical to the climate problem. Our aim is to provide a context for the design of modelling experiments and the presentation of model-based insights.

We begin with a brief description of the models themselves and a discussion of what we mean by 'climate'. Five sources of uncertainty are then described together with, where appropriate, the practical methods by which they may be explored in modelling experiments. Finally, we consider how we might communicate and increase our confidence in predictions using current models and consider the implications for experimental design and the balance of resources in climate modelling research.

2. Complex climate models

By the term 'complex climate models', we are referring to atmosphere/ocean global circulation models. These complex computer codes aim to mimic laboratories in other scientific disciplines; scientists use them to carry out experiments which are not possible in the real world. The atmospheric components have been developed as tools for weather forecasting (Cullen 1993). This application has benefited model development significantly by providing a cycle of model improvements and forecast system confirmation (Oreskes *et al.* 1994). Such confirmation only applies to processes with 'short' time scales on which we have many forecasts with corresponding (out-of-sample) observations. There are longer time-scale processes present, particularly in the ocean component, and additional long time-scale processes are being added (atmospheric chemistry, the carbon cycle, stratospheric dynamics, ice dynamics, etc.) as they develop into Earth System Models (ESMs). For these processes, and therefore for climate forecasting, there is no possibility of a true cycle of improvement and confirmation, the problem is always one of extrapolation and the life cycle of a model is significantly less than the lead time of interest (Smith 2002). Statements about future climate relate to a never before experienced state of the system; thus, it is impossible to either calibrate the model for the forecast regime of interest or confirm the usefulness of the forecasting process. Development and improvement of long time-scale processes are therefore reliant solely on tests of internal consistency and physical understanding of the processes involved, guided by information on past climatic states deduced from proxy data. Such data are inapplicable for calibration or confirmation as they are in-sample, having guided the development process. In any case, it would have only a weak role in terms of confirming the model as it applies to the system under different conditions. Failure to reproduce such observations usefully highlights model inadequacies and is valuable for model improvement, but success provides only a limited kind of confidence.

3. Climate

In the context of constant boundary conditions, and specifically no changes in atmospheric GHGs and therefore radiative forcing, weather is chaotic and climate may be taken as the distribution of states on some 'attractor' of weather. Under changing concentrations of atmospheric GHGs, the behaviour is not chaotic but pandemonium (Spiegel 1987). The distribution of possible weather, i.e. climate, has changed from the initial attractor but the distribution itself is in a transient state which may stabilize towards some other attractor when the forcing stabilizes at some other constant concentration.[1] Climate under climate change is still the distribution of possible weather but it cannot be evaluated in the real world (without access to many universes). It is well defined for a model (being the image of the initial attractor under the transient forcing) but its description would require a very large initial condition (IC) ensemble, which are suggested in future experimental designs. Since these are not currently available, we pragmatically take the distribution of ‘long’ time-scale (8-year) means over moderate size initial condition ensembles (ICEs), as indicative of climate. It is important to remember that climate change is a change in a distribution; most if not all decision support is sensitive to more than the mean of that distribution.

4. Sources of uncertainty

The interpretation of climate models to inform policy and decision support must consider at least five distinct sources of uncertainty. Forcing uncertainty captures those things in the future which are considered outside the climate system *per se*, yet affect it. Initial condition uncertainty captures our uncertainty in how to initialize the models in hand; what initial state, or ensemble of states, to integrate forward in time. Initial condition uncertainty is usefully divided into two camps depending on whether or not the details of today’s uncertainty in a variable are likely to influence the final distributions we estimate on our time scale of interest. Model imperfection describes the uncertainty resulting from our limited understanding of, and ability to simulate, the Earth’s climate. It is also usefully divided into two types: uncertainty and inadequacy. Model uncertainty captures the fact that we are uncertain as to what parameter values (or ensembles of parameter values) are likely to provide the most informative results; here, climate modelling has a further complication due to choices between parametrizations themselves, not just the values of each model parameter. Finally, model inadequacy captures the fact that we know *a priori*, there is no combination of parametrizations, parameter values and ICs which would accurately mimic all relevant aspects of the climate system. We know that, if nothing else, computational constraints prevent our models from any claim of near isomorphism with reality, whatever that phrase might mean. The five types of uncertainty are not independent and basic questions of identifiability and interpretation remain (Smith 2000). The design and interpretation of experiments in the face of these uncertainties are among the grand challenges of climate science today.

Interpretation is also complicated by the extrapolatory nature of the problem referred to earlier. Greenhouse gas scenario simulations relate to physical regimes (e.g. CO_2 levels) where we have no precise observations. In this sense, weather forecasting is closer to interpolation, as we have an archive of forecasts and

[1] The theory of nonlinear dynamical systems is founded on the long-time behaviours of parametrically static (or periodic) systems; definitions of ‘chaos’, ‘attractor’ and ‘mixing’, for example, cannot be applied to situations involving transient changes of parameter values. The development of a well-founded coherent jargon for this case would be of value in many fields of study.

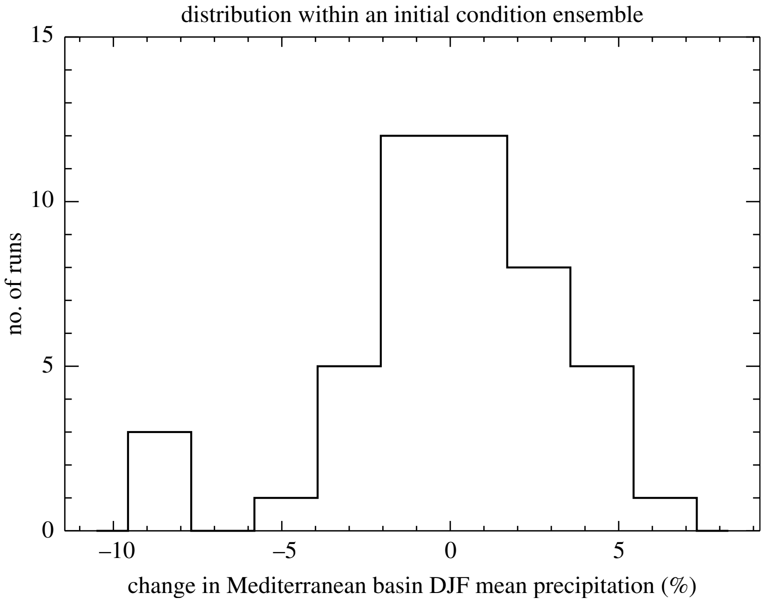

Figure 1. Distribution of the 8-year mean DJF Mediterranean basin precipitation across a 49 member IC ensemble, expressed as variations about the ensemble mean. Eight-year means were taken over years 7–15 in simulations with HadAM3.

corresponding observations which help us both to identify systematic inadequacies and to move back and forth between model states and the state of the atmosphere. There is no corresponding archive for climate.

Climate is a distribution. This distribution is sometimes taken to be a reflection of initial condition uncertainties (ICUs). If so, determining what the twenty-first century climate is, *conditioned on a particular* model, requires large ICEs. Exploration of ICU is important for interpretations of climate forecasts (Grimm *et al.* 2006). Figure 1 shows the distribution of 8-year mean December/January/February (DJF) precipitation over the Mediterranean basin in a 49 member ICE. The distribution is large; a range of approximately 17% at the regional level; much greater at the grid box level. There may well exist thresholds, or tipping points (Kemp 2005), which lie within this range of uncertainty. If so, the provision of a mean value is of little decision-support relevance. This is most likely to be important on regional scales where sustained changes over a decadal period could stimulate sustained changes in the region's climate. The response of glaciers is an obvious example. The ICUs and model imperfections are both an intrinsic part of a climate forecast which must be communicated to the users, a point we return to in the next section.

(*a*) *Forcing uncertainty*

One aspect of forcing uncertainty is explored using climate model simulations based on different scenarios of future concentrations of atmospheric GHGs. This is a familiar form of uncertainty: the likelihood of a given response depends in part on our actions. The likelihood of drowning is low in the shower, higher if we choose to swim in a shallow children's swimming pool, higher still in an adult pool and even higher along a beach with a strong undertow. Given that the anthropogenic GHG emissions are considered to be the most significant drivers of changes in climatic forcing in the twenty-first century (Cubasch *et al.* 2001; Stott *et al.* 2001), the future is therefore in 'our' control in the sense that we can choose if and where to swim. There is therefore no need to remove this uncertainty so long as reliable information can be given for the outcome of any particular choice. Our ability to give such information depends upon our ability to quantify the remaining sources of uncertainty. Such information should form the basis for making global-scale climate decisions. In practice, of course, socio-economic and psychological factors dominate this process (Muller 2002).

(*b*) *Initial condition uncertainty*

Given our time scale of interest, it can be useful to distinguish two forms of ICU depending on whether or not the model's future climate distribution would be better constrained by reducing the current level of uncertainty. Macroscopic ICU is found in state variables with *relatively 'large' slowly mixing scales* such that the predicted distribution is affected by our imprecise knowledge of the current state of the system. Microscopic ICU results from imprecise knowledge of *'small' rapidly mixing scales*. The consequences of these two forms of ICU give us the distribution which might be referred to as the climate of the twenty-first century conditioned on a particular model and forcing scenario. Microscopic ICU relates to distributions of model variability which will not be significantly changed by more accurate observations. Macroscopic ICU occurs in variables with longer time scales so that a better knowledge of these values can reduce the uncertainty at our target lead time; one example is details of the temperature–salinity structure in the oceans. Basin-scale variations in such quantities have been linked to decadal variations in regional climate in Western Europe and North America (Sutton & Hodson 2005). We should expect that different initial distributions of such variables would therefore produce different modelled climate distributions on our time scales of interest. The relevance of these climates could be increased by improved observations which reduce today's macroscopic ICU. Modelling experiments could identify further variables and regions of high dependency, and usefully direct observational campaigns.

By contrast, reducing microscopic ICU has no significant effect on the targeted climate distribution.[2] Its evolution is traced in weather forecasting where IC ensembles are used to produce probabilistic forecasts (Palmer 2000). Of course, the numerical weather prediction models are not perfect; they too are known to be fundamentally different from reality. Nevertheless, these ensembles can be interpreted as a source of information on which to build decision-support tools by exploiting the forecast archive and the daily addition of new, out-of-sample, observations (Palmer 2000; Smith 2006; Brocker & Smith in press). Climate forecasting cannot be approached in the same way. Unlike the weather case, we

[2] In a system which is mixing, ensembles drawn from two different initial distributions are distorted (mixed) to the extent that they appear as indistinguishable draws from the same final distribution. This can occur even if each ensemble member never 'forgets' its IC.

have no out-of-sample observations due to the time scales of interest and the lifetime of our models. Furthermore, the climate change problem is one of extrapolation so even given a probabilistic climate forecasting system which was somehow known to be reliable for the twentieth century, we would have no demonstrable reliability for the twenty-first century. Quantifying the impact of microscopic ICU is nevertheless of fundamental importance as it enables us to define the twenty-first century climate conditioned on our chosen model, forcing and perhaps some macroscopic ICs.

Microscopic ICU, like forcing uncertainty, is a familiar uncertainty. The proverbial parallel is rolling dice: assuming the die is truly random, it is simply a matter of chance what the outcome will be, although the probability of each face is thought to be well defined. In terms of predicting the model's response, that is the model's climate, this probability density function (PDF) can be deduced with a sufficiently large ICE based on a distribution of ICs. For climate time-scale simulations, small ICEs exist but they are insufficient to define the shape of the distribution for most variables and spatial scales.

Macroscopic ICU is less familiar and therefore less easily communicated. It describes the fact that there are some variables which inform us as to what distribution of future climate we are likely to sample. It can be compared with the time one has to wait for a bus at a bus stop. With no knowledge of the timetable, time of the day or day of the week, the situation may appear random. Knowledge of the timetable only allows some estimates to be made. Knowledge of the timetable and possession of a reliable watch would enable us to predict much better how long we have to wait on each occasion; the time of day being equivalent to a reduced macroscopic ICU, even if we do not know the day of the week which remains an unknown macroscopic ICU. Learning the day of the week would then reduce our uncertainty yet further.

(*c*) *Model imperfections*

We discuss two forms of model imperfection. 'Model uncertainty' describes the fact that climatic processes can be represented in models in different ways, e.g. different parameter values, different parametrization schemes and different resolutions (Smith 2000). It focuses us on determining the most useful parameter value(s) and model version(s) to study within the available model class, while accepting the impossibility of weather-like model calibration in climate studies. 'Model inadequacy' describes the fact that we know, even before running any simulations of the future, that even today's most complex, high-resolution models are unrealistic representations of many relevant aspects of the real-world system.

(i) *Model inadequacy*

It is a well known and widely discussed statistical problem that computer models of natural systems are inadequate (Oreskes *et al.* 1994; Kennedy & O'Hagan 2001; Bevan 2002; Chatfield 2002, 2003; Smith 2002) in the sense that they cannot be isomorphic to the real system. Nevertheless, such models might conceivably include or simulate to some degree of subjective adequacy, all the processes believed to be important in the system under study given the aims of the study. In cases where the reliability of the forecasting system cannot be confirmed, the ability of our models to reproduce the past observations in detail[3] gives us some hope that the model forecast may provide valuable guidance for the real world. Climate models fail this test.

First, they do not include many processes which are known to be important for climate change on decadal to centennial time scales, e.g. the carbon cycle, atmospheric and oceanic chemistry, and stratospheric circulation. Second, limitations due to grid resolution lead to some processes, which we would expect to result from the physics represented by the model being represented poorly, if at all. The models are known to have non-trivial shortcomings, examples include hurricanes, the diurnal cycle of tropical precipitation (Trenberth *et al.* 2003), many characteristics of El Niño-Southern Oscillation (ENSO) and the InterTropical Convergence Zone.

Model inadequacy is a serious barrier to the interpretation of model results for decision support, especially since, at present, there are no indications that the 'missing' processes would substantially ameliorate the climate change response, but an increasing number of plausible mechanisms which could make it worse than projected (Andreae *et al.* 2005; Walter *et al.* 2006). A frank discussion of which spatial and temporal scales today's best available information is thought to produce decision-relevant information would be of value. Consistency tests which may form a necessary condition for this determination have been discussed (Smith 2002).

Model inadequacy is a less familiar form of uncertainty; even statisticians often work pragmatically within an assumed structure (Chatfield 2002, 2003). It is particularly worrying in extrapolation problems, when we are moving into unfamiliar territory in which our past experience may not be a good guide for the future. If we have spent our whole life in Spain and then visit South Africa, without access to knowledge from people who have been there, we may anticipate there being different speed limits but nevertheless expect cars to drive on the right. We may be, perhaps fatally, surprised. Our 'model' of the way the world works might simply not include the possibility of such strange behaviour as driving on the left. This is a well-known philosophical problem referred to as Russell's (1946) chicken. In climate modelling, the situation may be worse; have we conceived of the possibility of driving on the left and then, for whatever reason, chosen to ignore it?

(ii) *Model uncertainty*

Model uncertainty includes uncertainty in the most relevant parameter values to be used in the model and can be quantified by ensemble experiments (Murphy *et al.* 2004; Stainforth *et al.* 2005). At the regional scale, it can be very large indeed (figure 2) and represents the impact of known uncertainties. Extending this from parameter values to parametrizations allows one to consider how to more usefully represent various processes within the model, and makes model uncertainty an extended form of the 'parameter uncertainty' of Kennedy & O'Hagan (2001). We suggest including elements which they might categorize as model inadequacy but only those which can be explored using different parametrization schemes, e.g. cloud formation and land surface effects in the current generation of models. Such parametrizations represent the processes' effects on the large scale and how they respond to the large-scale behaviour in the

[3] Ideally to shadow the past observations to within the observational uncertainties.

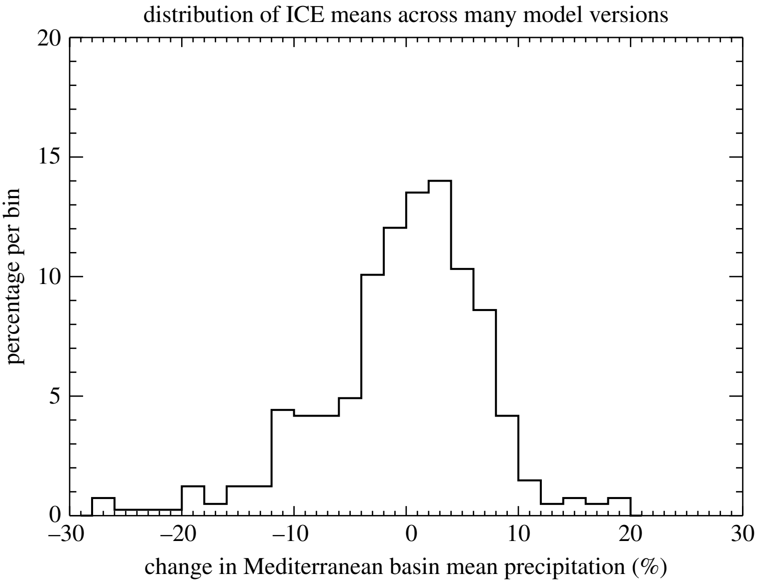

Figure 2. Distribution of the ICE mean change with doubling of atmospheric CO_2 concentrations, in 8-year mean Mediterranean basin DJF precipitation, in a grand ensemble (Stainforth *et al.* 2005). The change was taken as the difference in the mean over years 7–15, between a pre-industrial CO_2 and a double CO_2 simulation, for each version of HadSM3 (Williams *et al.* 2001) and set of ICs. Four hundred and eight model versions are included; each with an ICE of between 1 and 7 members.

model. They also contribute to model inadequacy when unable to reproduce feedbacks which occur at smaller scales or coupling across space and time scales, e.g. impacts of the diurnal cycle.

We can explore model uncertainty using ensembles of different models or model versions (Stainforth *et al.* 2005; referred to hereafter simply as models) which provide a range of plausible (state-of-the-art) simulations. Recent years have seen much activity in this direction, including the ensemble of opportunity used in the IPCC third assessment report (McAvaney *et al.* 2001), similar ensembles for the IPCC fourth assessment report (Meehl *et al.* submitted), the various model intercomparison projects (Covey *et al.* 2000), the grand ensemble of climate*prediction*.net (Stainforth *et al.* 2004, 2005) and the perturbed physics (Allen & Stainforth 2002) ensemble of the quantifying uncertainty in model predictions project (Murphy *et al.* 2004).

How should we interpret such ensembles in terms of information about future behaviour of the actual Earth system? The most straightforward interpretation is simply to present the range of behaviour in the variables of interest across different models (Stainforth *et al.* 2005, 2007). Each model gives a projected distribution; an evaluation of its climate. A grand ensemble provides a range of distributions; a range of ICE means (figure 2), a range of 95th centiles, etc. These are subject to a number of simple assumptions: (i) the forcing scenario explored, (ii) the degree of exploration of model and ICU, and (iii) the processes included and resolved. All analysis procedures will be subject to these assumptions, at least, unless a reliable physical constraint can be identified.

The frequency distributions across the ensemble of models may be valuable information for model development, but there is no reason to expect these distributions to relate to the probability of real-world behaviour. One might (or might not) argue for such a relation if the models were empirically adequate, but given nonlinear models with large systematic errors under current conditions, no connection has been even remotely established for relating the distribution of model states under altered conditions to decision-relevant probability distributions. The ensembles of opportunity are made up of state-of-the-art GCMs the results of which are inter-dependent due to the substantial collaborations between modelling centres, the publication of results in the literature, understanding based on the same text books and, in particular, the similarities of modern computer hardware across research centres (as all groups face similar constraints due to the limits of technology: for example, the upper bound on resolution is similar across models with no one running a global cloud resolving model at a resolution of a fraction of a kilometre on 100-year time scales). This reduces the confidence inspired by the agreement of results across a given generation of climate models, while increasing the utility of their disagreements for understanding processes in each individual model.

The shape of the frequency distribution of results in the perturbed parameter ensembles (figure 2) is governed by the characteristics of the base model and the choice of parameter values explored; at large spatial scales we might expect substantial parametric redundancy and thus a peak in the frequency distribution close to the behaviour of the base model (Stainforth *et al.* 2005). Even were we to achieve the impossible and have access to a comprehensive exploration of uncertainty in parameter space, the shape of various distributions extracted would reflect model constructs with no obvious relationship to the probability of real-world behaviour. Indeed, even the structure of parameter space is a subjective choice and has a first-order effect (Rougier *et al.* submitted). Physically, we can equally well use a parameter labelled ‘ice fall rate in clouds’ or its inverse (‘ice residence time in clouds’) and achieve identical simulations. Sampling uniform distributions under each of the two different labels however, yields completely different results. Under either choice of parameter label, the parameter is still merely a model parameter, which has at best a tenuous relation to the empirically well-defined velocity of any actual piece of ice (Smith 2006). It is unclear if the uncertainty in our empirical measurements has any parallel at all in terms of the plausible values of model parameter namesakes.

5. Interpretation of model results

Each model simulation provides a trajectory, a sample from the distribution that forms the future climate of the model at each point in time, under a given forcing scenario. Model inadequacies have substantial impacts on these projections on regional scales (and perhaps the global scale too). For many sources of inadequacy, the nonlinearity of the model suggests that we are unable to speculate on even the sign of that impact. Thus, how should we interpret and utilize these simulations? A pragmatic response is to acknowledge and highlight such unquantifiable uncertainties but to present results on the basis that they have zero effect on our analysis. The model simulations are therefore taken as possibilities for future real-world climate and as such of potential value to society, at least on variables and scales where the models agree in terms of their climate distributions (Smith 2002). But even best available information may be rationally judged quantitatively irrelevant for decision-support applications. Quantifying this relevance is a missing link in connecting modelling and user communities.

In the context of multi-model or perturbed physics ensembles, a wide range of possibilities are found. Wide though it is, we know that exploration of model and ICU has so far been limited, so we should expect this range to increase with further experiments. The range is therefore a lower bound on the maximum range of uncertainty; this range is −28 to 20% in figure 2.

Can we rule-out or weight the models in such ensembles? Much effort is currently focused on assigning models' weights based, to some extent, on their ability to reproduce recent climate (Murphy *et al.* 2004; Tebaldi *et al.* 2005; UKCIP08 2006). As long as all of our current models are far from being empirically adequate, we consider this to be futile. Relative to the real world, all models have effectively zero weight. Significantly non-zero weights may be obtained by inflating observational or model variability or by selecting a subset of variables or regions for the comparison. Both are misleading.[4] The former leads us to place more trust in a model whose mean response is substantially different (e.g. 5 standard errors) from observations than one whose mean response is very substantially different (e.g. 7 standard errors) from observations. A more constructive interpretation would be that neither is realistic for this variable and there is no meaning in giving the models weights based upon it. The latter assumes that interactions between variables, scales and locations are unimportant in this highly nonlinear complex system and therefore again overstates our understanding and confidence in a prediction. The lack of any ability to produce useful model weights, and to even define the space of possible models, rules out the possibility of producing meaningful PDFs for future climate based simply on combining the results from multi-model or perturbed physics ensembles; or emulators thereof. Models can, however, provide insight without being able to provide probabilities.

Even if we cannot assign relative weights to different models, we might nevertheless be able to provide pragmatic methods for deciding to include or exclude models from specific analyses. If the response of a model is to first order the result of a process which the model cannot capture by design, then we should assign that simulation zero weight; the model is outside its domain of applicability (Stainforth *et al.* 2005). We might also decide that all models whose simulations are 'substantially' worse than state-of-the-art models at some point in time, as assessed by their ability to simulate a wide range of observed variables, should be ruled out (Stainforth *et al.* 2005).

6. Confidence in model-based climate predictions

Computer experiments can be interpreted with two distinct aims: how to utilize today's model results and how to improve models; each aim suggests a different design for future experiments. Confidence is lower when the exploration of

[4] In fact, given a handful of imperfect models each with a million dimensional state-space, it should always be possible to find some metric in which the order of any two models is reversed.

uncertainty is less complete or when there is a lack of consistency in simulated response (consistency here would be in the distributions predicted by different models), while it is enhanced by the ability of the model to simulate observations of the processes of interest (see maximum consistency, below), and the agreement, in distribution, of models built on very different assumptions. In both cases, the possible impact of missing processes can be qualitatively discussed, at least in terms of their probable impact in the current climate. Confidence may also come from physical understanding of the processes involved. We have no hope of confirming the reliability of a predicting system, so confident statements about future climate will be more qualitative than one might wish and conditional on a number of significant assumptions. They may nevertheless by extremely valuable in guiding societal response to the climate change problem.

Today's ensembles give us a lower bound on the maximum range of uncertainty. This is an honest description of our current abilities and may still be useful as a guide to decision and policy makers. The statement implicitly conveys information on the confidence in the result and avoids over-interpretation which, if contradicted by the next generation of models, could undermine the credibility of climate science to inform policy, just at the point when more useful information might be becoming available. A method of utilizing such basic information to assess adaptation decisions is described in Stainforth *et al.* (2007).

Such an approach contains no information regarding the ability of the models to simulate the specific variables of interest. Failure to simulate the variables of interest under present-day climate, or recent changes therein reduces our confidence in the simulated changes being relevant to the real world. (For instance, model predictions of *changes in* the diurnal cycle of tropical precipitation should be taken with a pinch of salt given the lack of skill in the current simulations of this quantity.) Unfortunately, success does not imply reliability in the simulated future. Nevertheless, communicating such information may be useful. With a suitable grand ensemble, we can quantify the impact of ICU in the modelled variable and evaluate consistency with observations of the present day or the recent past. We can therefore replace figure 2—a histogram of modelled response in which only the maximum and minimum have any meaning—with a histogram showing the maximum consistency in each bin in response space (figure 3). This approach of comparative maximum consistency may be a useful way of communicating information about the quality as well as the response of model simulations in the ensemble. It could also be applied to derived variables of more direct interest to decision makers, for example potential crop yields or river flow statistics based on the simulations.

One might also use the ensembles to extract relationships which act as transfer functions between predicted variables and observed or better constrained quantities (Piani *et al.* 2005; Allen *et al.* 2006; Knutti *et al.* 2006). We illustrate this schematically in figure 4 which shows an ensemble-based relationship between a regional variable and global mean annual mean temperature ($T_{\rm g}$), a variable which has been most comprehensively studied, is susceptible to relatively simple analysis and in which we might arguably achieve the greatest confidence in prediction. Energy balance models can be combined with observations to infer global mean temperatures without involving global circulation models (Frame *et al.* 2005). In such studies, the assumptions of consistency with a small number of global-scale observations are at least clear. Where arguably robust relationships can be found between variables of interest

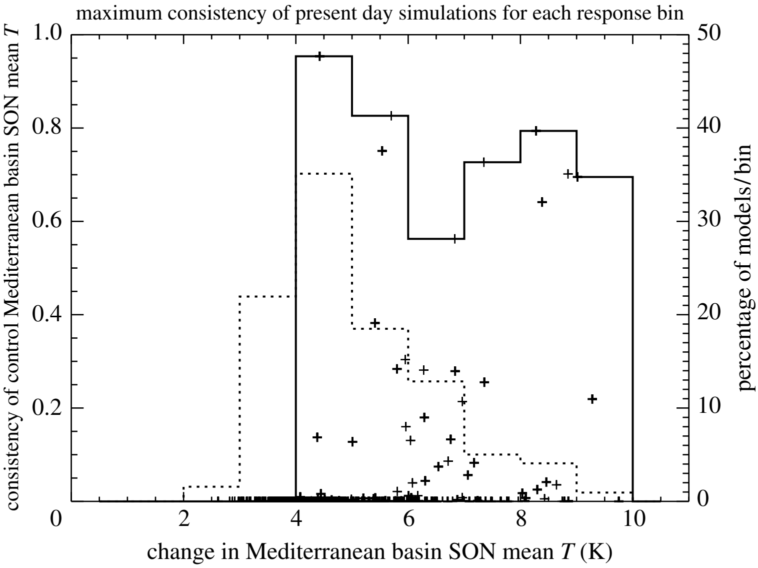

Figure 3. Illustration of a method to present comparative maximum consistency. The plot shows consistency with observations of the chosen model variable (8-year mean September/October/November mean Mediterranean basin temperature) for the range of responses in that variable to doubling of atmospheric CO_2 concentrations in a grand ensemble (Stainforth *et al.* 2005). The dotted line shows the frequency distribution of model versions (ICE means); right-hand axis. The points show the consistency for each model version; the solid line shows the maximum consistency in each bin. Consistency is taken as the probability that the ICE has the same mean as a set of observations: nine 10-year means over the period 1900–1990 from the IPCC Data Distribution Centre. It is calculated as the significance level in a student's t-distribution of a t-statistic for the two samples (observations and the ICE) defined by

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum_{i=0}^{N-1}(x_i - \bar{x})^2 + \sum_{j=0}^{M-1}(y_i - \bar{y})^2}{(N+M-2)}\left(\frac{1}{N} + \frac{1}{M}\right)}},$$

where x_i are the observations; y_i are the ICE members; N are the number of observation points (9); and M are the number of ICE members (between 2 and 7 inclusive). Three hundred and nineteen ICEs are included on the plot. Note: DJF data are not used in this plot because, in common with many variables, no ICE in this grand ensemble has a consistency above 2e–4.

and T_g, those relationships can be utilized to suggest probabilistic statements about our variable of interest, on the assumptions that this relationship reflects a physical mechanism and that it holds under an altered climate. A plausible, physically well-understood relationship would provide somewhat more confidence, although not mitigating the potential problems due to model inadequacy and limited exploration of uncertainties. It nevertheless suggests a somewhat higher level of consistency than simulations with a single model as it indicates a consistent response mechanism across multiple models. It seems clear that the use of large nonlinear models is necessary in climate science but in the analysis of their output we must clearly identify assumptions which imply simpler linear models would have sufficed, at least until we understand the physics of the linear relations our complex models have revealed to us.

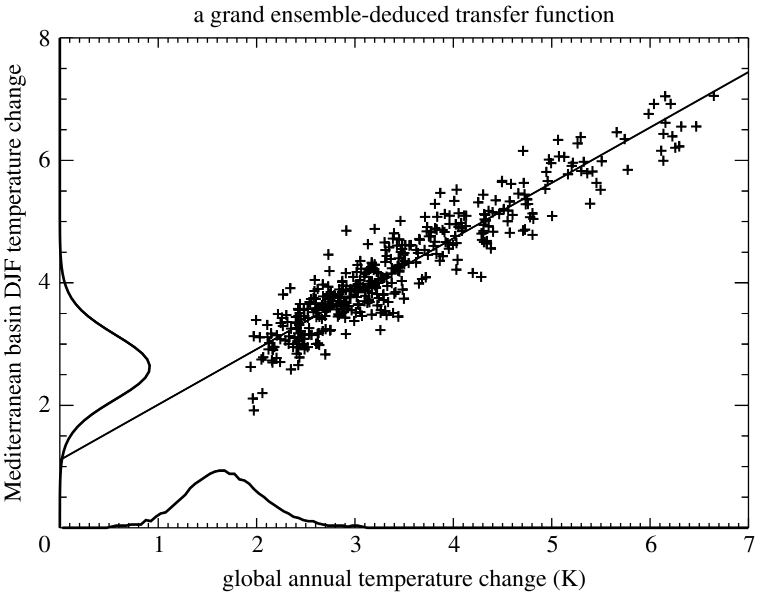

Figure 4. Schematic of the transfer method approach. ICE mean, 8-year mean Mediterranean basin DJF temperature change with 8-year mean global annual temperature change in a grand ensemble. A relationship exists in the ensemble and is used to create a distribution in the regional variable from a distribution in the global variable, in this case, the 2050s global temperature change under the A2 scenario from Stott & Kettleborough (2002).

7. Conclusions: the way forward

We have categorized various challenges in relating complex climate models to statements about the real world's future climate. The severity of model inadequacy suggests a more qualitative interpretation than one might wish. In particular, it is not at all clear that weighted combinations of results from today's complex climate models based on their ability to reproduce a set of observation can provide decision-relevant probabilities. Furthermore, they are liable to be misleading because the conclusions, usually in the form of PDFs, imply much greater confidence than the underlying assumptions justify; we know our current models are inadequate and we know many of the reasons why they are so. We have described several methods for presenting the results of large ensembles without assuming realism for any individual model, or indeed that any version of a model in the current class of models is close to realistic. These methods aim to increase our ability to communicate the appropriate degree of confidence in said results. Each model run is of value as it presents a 'what if' scenario from which we may learn about the model or the Earth system. Such insights can hold non-trivial value for decision making.

What then is the way forward for complex climate models and for their use in informing society about the changes we might expect in the Earth's climate through the twenty-first century? First, we must acknowledge that there are many areas for model improvement. Examples are the inclusion of a stratosphere, a carbon cycle, atmospheric/oceanic chemistry at some degree of complexity, ice-sheet dynamics, and realistic (i.e. statistically plausible equivalents of real-world behaviour) ENSO structures, land surface schemes (critical for exploration of regional feedbacks), diurnal cycles, hurricanes, ocean eddies and many others. We may be able to define a minimum acceptable resolution; suggestions of the order of 1 km in the horizontal have been discussed. In addition to improving the component parts, additional attention can be directed at the ability of the complete model to shadow the observations of the past (Smith 2000), noting when and how the model fails to shadow can suggest which components of the model are in most need of attention.

Models of such complexity, at high resolution and with suitable exploration of uncertainty are not going to be available soon. So what is the goal of model improvement activities? On the one hand, the answer is a tool for, and a part of, research on a wide range of climatic processes. A single complex climate model or ESM would prove ineffective here, as a comparison of climate distributions across models is a critical sanity check when extrapolating. On the other hand, the ultimate goal may be a tool for probabilistic predictions using techniques such as those proposed by Kennedy & O'Hagan (2001). In the meantime, they provide a range of possibilities which need to be considered. Revision of that range of possibilities based on more complex models will be useful; objective and robust methods to constrain them, even more so.

Perturbed physics experiments have demonstrated the large range of responses possible from the current class of models. There is no reason to expect this range to be substantially smaller in the next generation of models; increased physical realism may well increase it. Such uncertainty analyses are critical for the use of models to inform society. They are complemented by single simulations, or small ICEs, designed to explore the effects of specific factors, e.g. the carbon cycle. A two-pronged approach is required, involving model improvement and uncertainty assessments. One without the other is likely to be misleading. The balance of climate research effort is heavily skewed towards the former while the utilization of perturbed physics ensembles is heavily skewed towards the latter. Some rebalancing may benefit both research and its utilization in society.

Grand ensembles based on a range of base models are likely to expand uncertainty ranges and should be a priority in order to better understand the limits to our abilities today. New experimental designs need to include much greater exploration of ICU than those carried out to date. This is informative and valuable in itself, and is also important for the interpretation of large ensembles exploring model uncertainty. Experimental design can be focused to explore uncertainty in specific projections or aspects of the model; thus guiding model improvements as well as providing more generic uncertainty assessments.

There is much to be done but information from today's climate models is already useful. The range of possibilities highlighted for future climate at all scales clearly demonstrates the urgency for climate change mitigation measures and provides non-discountable ranges which can be used by the impacts community (e.g. Stern 2006). Most organizations are very familiar with uncertainty of many different kinds and even qualitative guidance can have substantial value in the design of robust adaptation strategies which minimize vulnerability to both climate variability and change. Accurate communication of the information we have is critical to providing valuable guidance to society.

We thank all participants in the climate*prediction*.net experiment and the many individuals who gave their time to make that project a reality. This work was supported by the Natural

Environment Research Council, the Tyndall Centre for Climate Change Research and the Framework 6 ENSEMBLES project.

References

Allen, M. R. & Stainforth, D. A. 2002 Towards objective probabilistic climate forecasting. *Nature* **419**, 228. (doi:10.1038/nature01092a)

Allen, M. R., Frame, D., Kettleborough, J. & Stainforth, D. A. 2006 Model error in weather and climate forecasting. In *Predictability of weather and climate* (eds T. Palmer & R. Hagedorn), ch. 15, pp. 391–427. Cambridge, UK: Cambridge University Press.

Andreae, M. O., Jones, C. D. & Cox, P. M. 2005 Strong present-day aerosol cooling implies a hot future. *Nature* **435**, 1187–1190. (doi:10.1038/nature03671)

Bevan, K. 2002 Towards a coherent philosophy for modelling the environment. *Proc. R. Soc. A* **458**, 2465–2484. (doi:10.1098/rspa.2002.0986)

Brocker, J. & Smith, L. A. In press. Increasing the reliability of reliability diagrams. *Weather Forecast.*

Chatfield, C. 2002 Confessions of a pragmatic statistician. *R. Stat. Soc. Ser. D—Stat.* **51**, 1–20. (doi:10.1111/1467-9884.00294)

Chatfield, C. 2003 *The analysis of time series: an introduction*, 6th edn. Boca Raton, FL: Chapman and Hall/CRC.

Covey, C., AchutaRao, K., Lambert, S. J. & Taylor, K. E. 2000 Intercomparison of present and future climates simulated by coupled ocean-atmosphere GCMs. *Technical report 66*—Program for climate model diagnosis and intercomparison. Livermore, CA: Lawrence Livermore Laboratory.

Cubasch, U. 2001 Climate change 2001. In *The science of climate change* (ed. J. T. Houghton), ch. 9, pp. 527–582. Cambridge, UK: Cambridge University Press.

Cullen, M. J. P. 1993 The unified forecast climate model. *Meteorol. Mag.* **122**, 81–94.

Frame, D. J. 2005 Constraining climate forecasts: the role of prior assumptions. *Geophys. Res. Lett.* **32**, L09702. (doi:10.1029/2004GL022241)

Giorgi, F. & Mearns, L. O. 2002 Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the ‘reliability ensemble averaging’ (REA) method. *J. Clim.* **15**, 1141–1158. (doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2)

Grimm, A. M., Sahai, A. K. & Ropelewski, C. F. 2006 Interdecadal variations in AGCM simulation skills. *J. Clim.* **19**, 3406–3419. (doi:10.1175/JCLI3803.1)

Hayhoe, H., Frumhoff, P., Schneider, S., Luers, A. & Field, C. 2006 Regional assessment of climate impacts on California under alternative emissions pathways—key findings and implications for stabilisation. In *Avoiding dangerous climate change,* ch. 24. Cambridge, UK: Cambridge University Press.

Kemp, M. 2005 Science in culture: inventing an icon. *Nature* **437**, 1238. (doi:10.1038/4371238a)

Kennedy, M. C. & O’Hagan, A. 2001 Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B—Stat. Methodol.* **63**, 425–450. (doi:10.1111/1467-9868.00294)

Knutti, R., Meehl, G., Allen, M. R. & Stainforth, D. A. 2006 Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Clim.* **19**, 4224–4233. (doi:10.1175/JCLI3865.1)

Lopez, A., Tebaldi, C., New, M., Stainforth, D., Allen, M. & Kettleborough, J. 2006 Two approaches to quantifying uncertainty in global temperature changes under different forcing scenarios. *J. Clim.* **19**, 4785–4796. (doi:10.1175/JCLI3895.1)

McAvaney, B. J. 2001 Climate change 2001. In *The science of climate change* (ed. J. T. Houghton), ch. 8, pp. 471–524. Cambridge, UK: Cambridge University Press.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J. & Taylor, K. E. Submitted. The global coupled climate multi-model dataset: a new era in climate change research.

Muller, B. 2002 *Equity in global climate change: the great divide.* Publication EV31. Oxford, UK: Oxford Institute for Energy Studies

Murphy, J. M. *et al.* 2004 Quantifying uncertainties in climate change from a large ensemble of general circulation model predictions. *Nature* **430**, 768–772. (doi:10.1038/nature02771)

O'Hagan, A. & Oakley, J. E. 2004 Probability is perfect, but we can't elicit it perfectly. *Rel. Eng. Syst. Safety* **85**, 239–248. (doi:10.1016/j.ress.2004.03.014)

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994 Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641–646. (doi:10.1126/science.263.5147.641)

Palmer, T. 2000 Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**, 71–116. (doi:10.1088/0034-4885/63/2/201)

Piani, C., Frame, D. J., Stainforth, D. A. & Allen, M. R. 2005 Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.* **32**, L23825. (doi:10.1029/2005GL024452)

Rougier, J. *et al.* Submitted. Emulating the sensitivity of the HadSM3 climate model using ensembles from different but related climate models.

Russell, B. 1946 *The problems of philosophy*, p. 63, 2nd edn. Oxford, UK: Oxford University Press.

Smith, L. A. 2000 Disentangling uncertainty and error: on the predictability of nonlinear systems. In *Nonlinear dynamics and statistics* (ed. A. I. Mees), pp. 31–64. Boston, MA: Birkhauser.

Smith, L. A. 2002 What might we learn from climate forecasts? *Proc. Natl Acad. Sci. USA* **99**, 2487–2492. (doi:10.1073/pnas.012580599)

Smith, L. A. 2006 Predictability past predictability present. In *Predictability of weather and climate* (eds T. Palmer & R. Hagedorn), ch. 9, pp. 217–250. Cambridge, UK: Cambridge University Press.

Spiegel, E. A. 1987 Chaos—a mixed metaphor for turbulence. *Proc. R. Soc. A* **413**, 87–95. (doi:10.1098/rspa.1987.0102)

Stainforth, D. A., Allen, M. R., Frame, D., Kettleborough, J., Christensen, C., Aina, T. & Collins, M. 2004 Climateprediction.net: a global community for research in climate physics. In *Environmental online communication* (ed. A. Scharl), pp. 101–112. London, UK: Springer.

Stainforth, D. A. *et al.* 2005 Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406. (doi:10.1038/nature03301)

Stainforth, D. A., Downing, T. E., Washington, R., Lopez, A. & New, M. 2007 Issues in the interpretation of climate model ensembles to inform decisions. *Phil. Trans. R. Soc. A* **365**, 2163–2177. (doi:10.1098/rsta.2007.2073)

Stern, N. 2006 Stern review on the economics of climate change. HM Treasury.

Stott, P. A. & Kettleborough, J. A. 2002 Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726. (doi:10.1038/416723a)

Stott, P. A. *et al.* 2001 Attribution of twentieth century temperature change to natural and anthropogenic causes. *Clim. Dyn.* **17**, 1–21. (doi:10.1007/PL00007924)

Stott, P. A., Kettleborough, J. A. & Allen, M. R. 2006 Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.* **33**, L02708. (doi:10.1029/2005GL024423)

Sutton, R. T. & Hodson, D. L. R. 2005 Atlantic Ocean forcing of North American and European summer climate. *Science* **309**, 115–118. (doi:10.1126/science.1109496)

Tebaldi, C., Smith, R. L., Nychka, D. & Mearns, L. O. 2005 Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J. Clim.* **18**, 1524–1540. (doi:10.1175/JCLI3363.1)

Trenberth, K. E., Dai, A., Rasmussen, R. M. & Parsons, D. B. 2003 The changing character of precipitation. *Bull. Am. Meteorol. Soc.* **84**, 1205–1217. (doi:10.1175/BAMS-84-9-1205)

UKCIP02 2002 UK climate impacts programme scenarios. See http://www.ukcip.org.uk/scenarios/.

UKCIP08 2006 Expressed preferences for the next package of UK climate change information. UKCIP report. See http://www.ukcip.org.uk/scenarios/ukcip08/documents/User_Consultation_report_v5.pdf.

Walter, K. M., Zimov, S. A., Chanton, J. P., Verbyla, D. & Chapin III, F. S. 2006 Methane bubbling from Siberian thaw lakes as a positive feedback to climate warming. *Nature* **443**, 71–75. (doi:10.1038/nature05040)

Williams, K. D., Senior, C. A. & Mitchell, J. F. B. 2001 Transient climate change in the Hadley centre models: the role of physical processes. *J. Clim.* **14**, 2659–2674. (doi:10.1175/1520-0442(2001)014<2659:TCCITH>2.0.CO;2)