PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2008) **366**, 4647–4664
doi:10.1098/rsta.2008.0169
Published online 25 September 2008

Should we believe model predictions of future climate change?

By Reto Knutti*

Institute for Atmospheric and Climate Science, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland

Predictions of future climate are based on elaborate numerical computer models. As computational capacity increases and better observations become available, one would expect the model predictions to become more reliable. However, are they really improving, and how do we know? This paper discusses how current climate models are evaluated, why and where scientists have confidence in their models, how uncertainty in predictions can be quantified, and why models often tend to converge on what we observe but not on what we predict. Furthermore, it outlines some strategies on how the climate modelling community may overcome some of the current deficiencies in the attempt to provide useful information to the public and policy-makers.

Keywords: climate model; uncertainty; probability; climate prediction; future climate change

1. Introduction

There is clear and widespread evidence that the climate on Earth is changing rapidly. Human influence is very likely responsible for many of the observed large-scale trends of warming, and these are almost certain to continue in the near future (IPCC 2007). However, in order to make decisions for mitigation and local adaptation, we need to know how the climate will change locally, and the extent to which heavy rainfall events, sea ice, permafrost, heat waves and many other aspects of the climate are likely to change. To predict these changes, we inevitably have to rely on complex numerical computer models. To what extent should we trust the numbers that come out of our models?

When we look at a weather forecast from a numerical model that indicates a 70 per cent chance of rain, we decide whether to take an umbrella or not. Our decision is based on experience from similar situations in the past, rather than on an understanding of the underlying model. When we look at climate projections for the next century, in fact, from similar models, we are torn between believing, questioning and ignoring them. Why is it so difficult to communicate what we know and what is uncertain about future climate change? Why are climate model projections uncertain anyway? How can we be sure that a model that performs well in simulating past or present climate will be reliable in simulating future climate?

*reto.knutti@env.ethz.ch

One contribution of 10 to a Triennial Issue ‘Earth science’.

The answers to these questions depend on the time scale, the spatial scale, the variable (e.g. temperature, precipitation, sea ice, or sea-level rise) and the statistic (i.e. the mean of a quantity, its trend, the change in variability or extremes) in which we are interested. In this paper, rather than providing a final answer (which is unlikely to exist) or attempting to put a skill score on a climate model, I try to outline how climate scientists think about their models, how they develop and improve them, and how they interpret their results. I discuss sources of model bias and uncertainty and ways to evaluate models, and I try to explain why it is so difficult to quantify model performance. A few ideas are given at the end as to how the need for 'near operational climate projections' may require us to rethink the way decisions are made in the model development and evaluation process.

2. What is a climate model, and why do we need more than one?

Climate models consist of a set of equations that are discretized on a grid and solved numerically on a large computer. Some equations are derived from first principles (e.g. equations of motion, and conservation of energy, mass and angular momentum), but many processes have to be parametrized in a simplified form. For example, there is no known equation to describe the effect of a growing tree on the climate, yet trees modify the surface properties of the Earth and thus affect the exchange of energy, water and carbon with the atmosphere on local to global scales. Hence, the effect of the tree is parametrized in terms of environmental conditions (type of tree, water availability, light availability, temperature, nutrients, competing plants, etc.). For the parts of the model governed by fundamental equations (e.g. the equations of motion), increased computational capacity and thus finer resolution will improve the simulation. For empirical relationships where there is no fundamental underlying law (such as the effect of the tree on the land surface), the limiting factor is probably our understanding rather than the computational capacity.

A variety of models have been developed to study different aspects of the climate system. There is no single best model, but rather a pool of models or model components that are combined to study specific questions. The decisions as to what parts of the system are modelled explicitly and what are fixed or externally prescribed is guided by the question of interest as well as by practical considerations such as computational capacity. The most complicated model is often not the easiest to interpret and may present computational challenges. Therefore, a hierarchy or spectrum of models from one-dimensional energy balance models through models of intermediate complexity to fully coupled atmosphere–ocean general circulation models (AOGCMs) describing the atmosphere, ocean, sea ice, land and possibly chemistry, the carbon and nutrient cycles and ice sheets has been developed (Claussen *et al.* 2002). If we accept the need for different models for different questions, then why do we also have multiple models of similar complexity to study the same question? While these models may be seen as incompatible in the sense that they involve conflicting assumptions, the more common view is that they all are credible approximations to the description of the climate system given our limited understanding, the lack of complete observations, and the simplifications that need to be made due to computational constraints (Parker 2006). While model development is in some

sense competitive, the existing set of AOGCMs for example is mostly seen as a family of coexisting models that sample to some extent the uncertainty in describing the system.

Models need to serve multiple purposes. Model development can be driven by curiosity and the attempt to understand and quantify processes. Simpler models and idealized set-ups are often easier to interpret in that context. However, the urgency of climate change has partly shifted the focus towards making predictions and projections (projections indicate that the outcome is conditional on some hypothesis about the future, e.g. the rate of CO_2 emissions in an externally prescribed economic scenario). For projections, the aim is to provide robust and useful information to assist policy-makers in the decision-making process. To satisfy the desire for ever more detailed information, the strategy was, and still often is, to take the most comprehensive model that is affordable, run a few simulations on the largest computer available at the highest resolution possible and then struggle to make sense of the results.

3. Why are projections from climate models uncertain?

Projections from models are inherently uncertain, because a model can never fully describe the system that it attempts to specify. All models are always imperfect to some extent (e.g. Oreskes *et al.* 1994; Kennedy & O'Hagan 2001; Smith 2002; McWilliams 2007). Uncertainty in model projections arises from boundary and initial condition uncertainty, our incomplete theoretical understanding of the system, parameter uncertainty and the fact that our models are imperfect (see Stainforth *et al.* (2007) for a review). Boundary conditions (e.g. CO_2 concentrations from an economic scenario) are prescribed externally to the model, and results are often interpreted simply as conditional on the boundary conditions. If the time scale of the projection is much larger than the memory time scale of the system, and the system is chaotic, then the initial condition problem is circumvented by running multiple ensemble members (simulations with the same model, parameters, boundary conditions and scenario, but slightly different initial conditions) or by averaging over longer time periods (multiple ensemble members are usually preferred but more expensive to compute). Better knowledge of initial conditions may help in improving forecasts on shorter time scales, e.g. seasonal forecasts or predictions of the El Niño Southern Oscillation (ENSO), but for most of the climate change projections, uncertainties in the model structure and parameters dominate.

For some processes, we simply do not know enough to be able to simulate a process with a model. One such example is how exactly clouds form. To sample these uncertainties, multiple empirical (but typically only few) parametrizations of the same process exist. Parameter uncertainty arises from the fact that the parameter values used in these parametrizations are not always well constrained by the observational evidence. Finally, all current climate models are known to be empirically inadequate in the sense that no set of parameters can always fit the observations within their uncertainty (e.g. Sanderson *et al.* 2008). This is often referred to as structural error. The choices made in the model structure (e.g. the numerical schemes used, the decision for a finite volume versus a spectral representation of the atmosphere, or whether vegetation cover is fixed or

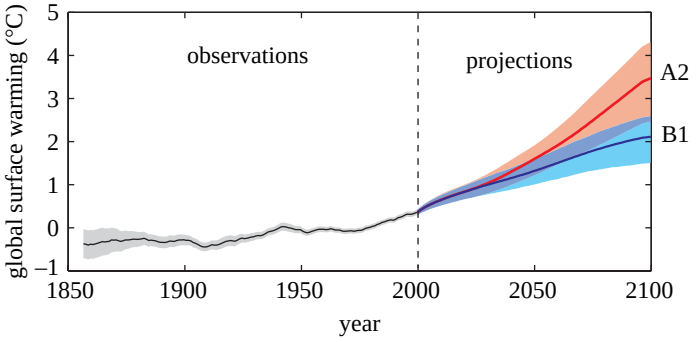

Figure 1. Observed surface warming (black) and its uncertainty (grey) (Jones *et al.* 1999) and 5–95% confidence range for simulated future global temperature change for two non-intervention scenarios (SRES B1 and A2, colours). The uncertainty in projections is constrained by the observed global surface warming and ocean heat uptake. Modified from Knutti *et al.* (2002).

variable), as well as computational constraints limiting the resolution of the model, introduce limitations in the model that are often persistent for any set of parameters chosen.

As the spread of a few models or initial condition ensembles are not good measures of uncertainty, other attempts to quantify uncertainty and probability in projections using, for example, Bayesian methods (Tebaldi *et al.* 2005; Furrer *et al.* 2007), perturbed physics ensembles (Forest *et al.* 2002; Knutti *et al.* 2002) and pattern scaling arguments (Stott & Kettleborough 2002) have seen a boost in recent years, but are still in their infancy (see Tebaldi & Knutti (2007) for a review). An example of such a probabilistic projection for two scenarios (SRES B1 and A2; Nakicenovic & Swart 2000) is shown in figure 1 based on the Bern2.5D climate model of intermediate complexity (Knutti *et al.* 2002). It shows 5–95% confidence ranges for global temperature projections, constrained by the observed surface warming (Jones *et al.* 1999) and ocean heat uptake (Levitus *et al.* 2000), using the additional condition that climate sensitivity is less than 4.5°C (see Knutti *et al.* (2002) for details). Results from different studies are usually consistent within their stated uncertainties (Knutti *et al.* 2008), but the structural differences between studies are rather large.

Uncertainty naturally depends on the spatial and temporal scales and variables considered. Many components of the climate system are chaotic, but on large spatial scales and on decadal or longer time scales are believed to be predictable. The above-mentioned sources of uncertainty will contribute differently to various questions, and some quantities are inherently more difficult to predict than others. Ideally, the assumptions, conditions, uncertainties and underlying framework of models would need to be communicated to those making decisions, for them to be able to evaluate the relevance of the information provided by the models, and to make informed decisions. This often turns out to be too complex, the technical or time consuming, so it is the scientist's duty to provide guidance on the interpretation of the model results.

The question asked at the outset of whether we should believe anything that our models predict about future climate is related to how well we can quantify the uncertainty in model projections. To quantify uncertainty, one needs to decide whether a model is 'credible', 'plausible', or consistent with observations (given some known limitations of the model).

4. What is a good model?

Any model that describes an open system cannot strictly be verified, i.e. proven to be true, nor can it be validated in the sense of being shown to accurately represent—both at present and for all future times—the processes responsible for the observed behaviour of the real system (Oreskes *et al.* 1994). However, in certain cases, we may be able to provide excellent evidence regarding the processes that are at work in a system at a given time. The process of testing a model and comparing model results with observations is usually referred to as model evaluation, whereas selecting values for unknown or uncertain parameters is known as calibration. The latter is also referred to as tuning, but that word has a negative undertone of being dishonest and adjusting parameters to get the right effect for the wrong reason, when in fact calibration involves estimating parameters based on either limited observations or physical understanding. So the best we can hope for is to demonstrate that the model does not violate our theoretical understanding of the system and that it is consistent with the available data within the observational uncertainty. For climate projections the situation is more difficult. Model calibration is strictly impossible in this case, as projections of future climate change relate to a state never observed before. Making a projection for the climate in the year 2100 and waiting a century for the data to evaluate the projection is unfeasible; also, a single realization of the climate may not tell us much anyway. The problem is that the life cycle of a model is much shorter than the time scale over which a prediction can be evaluated by observations (Smith 2002).

We, therefore, need to assume that the equations, parametrizations and assumptions built into the model can be extrapolated beyond the range of where they are evaluated. This is obvious for parts of the model (e.g. conservation of mass or energy), but less clear in the case of parametrizations that are empirically derived from observations mostly covering the rather narrow climate regime of the past century.

Further complications arise because the datasets used in the model evaluation process are mostly the same as those used for calibration, which may result in circular reasoning. For example, the ocean diffusivity is chosen to give good agreement with the distribution of ocean temperature, and the same temperature fields are used later for model evaluation. There is a danger that model–data fit will be artificially inflated, giving a poor indication of how accurate the models will be when it comes to new, out-of-sample predictions. The model calibration is often non-unique, such that different sets of model parameters may reasonably reproduce the observations (e.g. Knutti *et al.* 2002; Stainforth *et al.* 2005). The obvious idea is to throw more data at the problem, in the hope that the parameters will be constrained if the amount of data for the evaluation is much larger than the number of adjustable parameters. However, more detailed observations often require a more complex model, and the more comprehensive a model, the more difficult it is to evaluate and understand it (Oreskes 2003). In addition, more constraints usually make the model imperfection more obvious. The model may be calibrated to one dataset, but it may not be consistent with datasets of several quantities simultaneously (e.g. Sanderson *et al.* 2008).

Climate models make projections of many different quantities, so whether a model is 'good' or 'bad' depends on the question of interest. For numerical weather prediction, for example, skill is relatively well defined because forecasts can be verified on a daily basis, which makes it possible to derive robust statistics of model skill (although there are different skill scores as well). For a climate model, it is more difficult to define a unique overall figure of merit, metric or skill score for long-term projections. Each model tends to simulate some aspects of the climate system well and others not so well, and each model has its own set of strengths and weaknesses. The situation is worse because skill has to be defined based on the simulation of processes, the past or the present climate, rather than the quantity of interest, the future, for which no observations exist.

If model and data do not agree, the scientist should be worried (assuming the model was designed to reproduce the data), because it implies that the model (or the data) is in error. The opposite, however, is not true. Agreement between model and data does not imply that the modelling assumptions accurately describe the processes producing the observed climate system behaviour; it merely indicates that the model is one (of maybe several) that is plausible, meaning that it is empirically adequate. All AOGCMs, for example, reproduce the observed surface warming rather well (Hegerl *et al.* 2007; Knutti 2008), and yet they have different feedbacks that lead to different projections for future warming, so the agreement over the past is not *per se* a guarantee for a reliable prediction. A lack of disagreement in repeated experiments and applications means that the model is more likely to be adequate and useful to infer some prediction from it, at least within the range of applications or parameters where it has been evaluated (Oreskes *et al.* 1994). Model performance is also relative; a good model is often meant to be good relative to other models of similar type.

A pragmatic approach is to accept that the observed data tell us only about internal consistency between model and data. We do not need a perfect model, just one that serves the purpose. An aeroplane, for example, can be constructed with the help of numerical models that are not able to properly simulate turbulent flow. Economic models are used for risk management, knowing that it is impossible to properly describe human behaviour in mathematical terms. A model only needs to be empirically adequate to address a particular question. In the absence of better alternatives, one may even use a 'model' that is not known to be adequate but that is likely to be better than, say, random guessing. For example, a broker invests money into a company that develops a product. He has no quantitative, formal model known to have skill in guiding such investments; he makes judgements based on past performance of the company (implicitly his model) in different situations. His decisions are essentially an extrapolation, but are still likely to be better (at least that is what he believes) than random guessing.

A collection of models can be used to characterize the uncertainty in projections (Tebaldi & Knutti 2007). On the largest scale, the models can be seen as different plausible approximations to the real world given the uncertainty in our understanding of the climate system (Parker 2006), the limited observations and the constraints in computational capacity. Models may make conflicting assumptions on smaller scales, i.e. two parametrizations of ocean mixing may

strictly be seen as inconsistent in the structure of the equations, but both would agree with observations within some range, which typically is the sum of the uncertainty of the observations and the model error. A coarse resolution ocean model may parametrize mixing by eddies as diffusion, while an eddy resolving model will explicitly simulate eddies and the transport they induce. Neither of them perfectly simulates in complete detail the process of mixing by eddies in the real world, but both are plausibly considered useful representations, given the goals of the modelling study.

The inevitable question with a collection of models is whether some are more plausible than others. Model projections have been weighted explicitly by how well they reproduce the present-day climate (e.g. Tebaldi *et al.* 2005). The critical assumptions in most of these Bayesian studies attempting to generate probabilistic information (probability density functions, PDFs) are that all models are independent, distributed around a perfect model of the climate system, and sample the range of uncertainty. Unfortunately, none of these assumptions are strictly true for the current set of AOGCMs. The models are not independent (Jun *et al.* 2008*a*) and do not sample the range of uncertainties even on the largest scales (Knutti *et al.* 2008). Because each institute usually develops only one or two models, the current set of AOGCMs should be interpreted as the collection of the best model that each group could build, each one of them carefully calibrated to the same (possibly biased) observations, rather than an ensemble trying to explore uncertainties. Therefore, more models (parameter perturbations to existing models or structurally different models with similar performance) should not necessarily improve our confidence, because the structural uncertainty or model imperfection is not addressed by sampling from the same class of models (see Tebaldi & Knutti (2007) for a detailed discussion).

Some scientists argue that we cannot attach weights to models, produce meaningful PDFs or even define the space of plausible models, because all models have essentially zero weight (Stainforth *et al.* 2007). If the requirement is to match all observations, then this may strictly be true. However, one can argue that the models do not have zero weight in predicting future global temperature, because they do not need to reproduce all observations. In addition, likelihoods are already inferred implicitly all the time. Newer models are used in reports of the Intergovernmental Panel on Climate Change (IPCC) to make projections (Meehl *et al.* 2007*b*), whereas older models are discarded (i.e. get zero weight). Individual studies select subsets of models based on subjective criteria (van Oldenborgh *et al.* 2005). The end-user always assumes that the newest models provide the best information, and that the model spread provides some estimate of uncertainty. PDFs may indeed be questionable and imply too much certainty, but I argue that the problem lies more in how to interpret them than whether they should be constructed in the first place. PDFs should be seen as an indication of what outcomes may be more plausible than others, or as a way to communicate uncertainty, rather than as a strict mathematical representation of it. The PDFs themselves are uncertain or 'fuzzy' (Kriegler & Held 2005; Knutti *et al.* 2008), and obviously conditional on the model, the observations and the statistical method used, and on the sources of uncertainties considered.

5. Should we have confidence in our model projections?

The recent World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl *et al.* 2007*a*), a set of coordinated simulations with over 20 AOGCMs, has led to an unprecedented effort with literally hundreds of studies evaluating and comparing the current AOGCMs (Räisänen 2007; Randall *et al.* 2007 and references therein; Gleckler *et al.* 2008; Reichler & Kim 2008) (see also http://www-pcmdi.llnl.gov/ipcc/subproject_publications.php). All models simulate the present-day temperature and (to some degree) precipitation well on large scales (Randall *et al.* 2007), and simulated trend patterns are consistent with observations (Hegerl *et al.* 2007) if models are forced with all radiative forcings. Projected future warming patterns are robust (Meehl *et al.* 2007*b*), but global temperature change is uncertain by approximately 50 per cent (Knutti *et al.* 2008) owing to carbon cycle uncertainties (Friedlingstein *et al.* 2006) and models differing in their feedbacks (Bony *et al.* 2006). Short-term projections are better constrained by the observed warming than long-term projections (Knutti *et al.* 2002; Stott & Kettleborough 2002). Models project changes in precipitation, extreme events (Tebaldi *et al.* 2006) and many other aspects of the climate system that are consistent with our understanding, but agreement between models deteriorates as one moves from continental to regional to local (i.e. grid point) scales.

The average of multiple models outperforms any single model when compared with observations (Gleckler *et al.* 2008; Reichler & Kim 2008), indicating that part of the biases are random. However, models are often based on similar assumptions, such that structural errors are correlated (Tebaldi & Knutti 2007; Jun *et al.* 2008*a*). Some models also agree better with observations than others; therefore, naive averaging is unlikely to be the best option. In figure 2 the 'model performance mountain' is shown, an attempt to visualize the CMIP3 model landscape and performance. The horizontal position of each model is determined by a 'classical multidimensional scaling' analysis (principal coordinate analysis) of a distance matrix, where the distance metric between two models is the root-mean-square difference of the simulated 1980–1999 mean winter and summer surface temperature maps. The multidimensional scaling (Cox & Cox 2001) finds the best representation of the models in a two-dimensional space, such that models being 'close' in their distance metric tend to be close on the map (for a matrix of distances between European cities, the algorithm would produce a map of Europe). The vertical coordinate is the root-mean-square difference of the simulated 1980–1999 surface temperature compared with the ERA40 reanalysis dataset (located on top of the mountain). The better the agreement, the higher the model's elevation on the mountain. The figure illustrates that models 'climb the mountain from different sides', but that models developed by the same group (same symbols) tend to be near each other, since they show similar bias patterns (e.g. the three NASA GISS models 8–10 are all on one side of the mountain). The multi-model average (blue) is slightly better than some of the best models, and the average of the five best models (green) is even better, but they are not as good as one would expect if all biases were random. The figure will, of course, look different for each variable and should, therefore, be seen as one depiction of model similarity rather than an

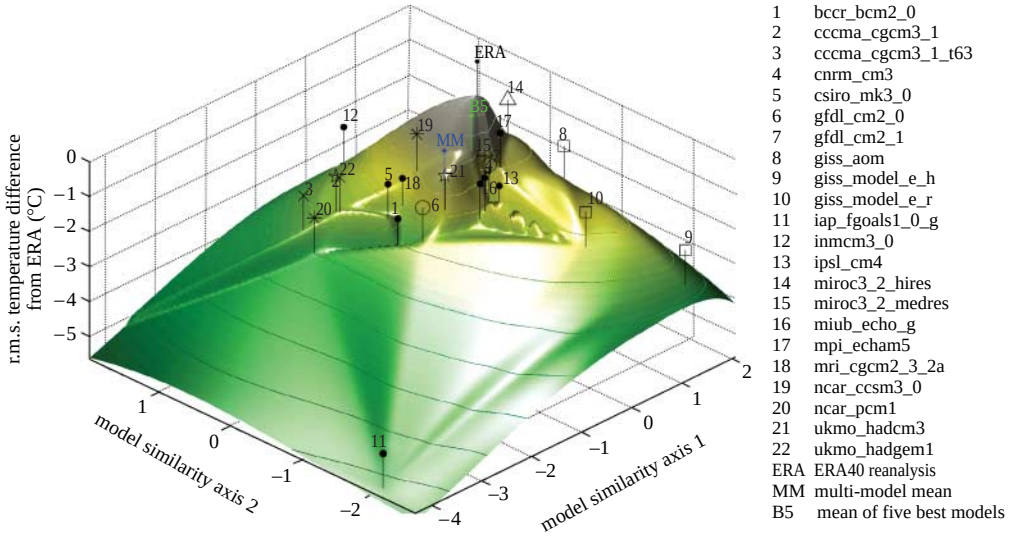

Figure 2. Climbing the model performance mountain. The horizontal distance between two models is an approximation to their similarity in simulating the spatial distribution of present-day surface temperature. The horizontal coordinates are dimensionless and have no physical meaning. The elevation of a model is the r.m.s. difference of climatological surface temperature from observations; perfect agreement with reanalysis is on top of the mountain (see text for details). Models from the same institution (same symbols except dots, which mark single models) tend to be near each other, because they have similar bias patterns. The multi-model mean (MM, blue) and the mean of the five best models (B5, green) perform better than single models and are closer to the top of the mountain, but in terms of elevation still far away from it.

objective quantification of skill in some overall sense. For a more detailed discussion of local correlation of model biases and visualizations using multidimensional scaling, see Jun *et al.* (2008*b*).

Confidence in climate models comes from different lines of evidence. First, models are based on physical principles such as conservation of energy, mass and angular momentum. Second, models reproduce the mean state and variability in many variables reasonably well, and continue to improve in simulating smaller-scale features (Räisänen 2007; Randall *et al.* 2007; Gleckler *et al.* 2008; Reichler & Kim 2008). Progress has been made in understanding climate feedbacks and intermodel differences (Bony *et al.* 2006). Model calibration from observations is unavoidable for certain subcomponents of the models, but the amount of data is large and the number of model parameters is small, so the calibration should not be mistaken as a statistical curve fitting exercise. Third, models reproduce observed global trends and patterns in many variables (Barnett *et al.* 2005; Hegerl *et al.* 2007). Fourth, models are tested on more distant past climate states (e.g. Jansen *et al.* 2007), which provides a useful and relatively independent evaluation, although uncertainties in evidence of the distant past are larger and proxy data may not directly be used to test the model. Also, some feedbacks or model assumptions may no longer be fully valid under different boundary conditions. Fifth, multiple models agree on large scales, which is implicitly or explicitly interpreted as increasing our confidence (Tebaldi *et al.* 2005). Sixth, projections from newer models are consistent with older

ones (e.g. for temperature patterns and trends; Cubasch *et al.* 2001; Meehl *et al.* 2007*b*), indicating a certain robustness. Some aspects of the climate system (e.g. thresholds) are likely to be structurally unstable (such that arbitrarily small changes in the model structure, initial conditions or parameters cause changes in the solution that are not small, as in the case for numerical weather prediction) (Knutti & Stocker 2002; Smith 2002; McWilliams 2007). However, the consistency across models suggests that, for many large-scale features of the climate, the projected changes are 'structurally stable' (McWilliams 2007). Finally, confidence comes from the fact that we can understand results in terms of processes. The model results we trust most are those that we can understand the best, and relate them to simpler models, conceptual or theoretical frameworks.

George Box is credited with the quote that 'All models are wrong, but some are useful' (Box 1979). Indeed, all climate models are known to be imperfect to some degree, but they can still help us to understand the things we observe or simulate and to test hypotheses. Simplified models, e.g. simulations of an aquaplanet (in simple words, an atmosphere with a zonally uniform ocean underneath and no continents) (e.g. Medeiros *et al.* 2008) with idealized boundary conditions, can be particularly powerful in that context, as processes can be traced through a hierarchy of models, and the behaviour of a model can be studied as different model components are added or removed (Held 2005).

An interesting recent development is that models predict changes that are observed later (e.g. that the ocean should have warmed over the last century (Levitus *et al.* 2000)), or that discrepancies between models and data are found to be a problem of the data (Santer *et al.* 2003; Thompson *et al.* 2008), thus abandoning the circle of evaluating the models with in-sample data.

The downside is that the model spread into the future is often not decreasing, e.g. the spread of climate sensitivity or transient climate response in CMIP3 is almost the same as in the previous model generation. In addition, many problems seem to be similar across families of models, because models make similar assumptions. The structural error of the current models is, therefore, significant (Tebaldi & Knutti 2007; Sanderson *et al.* 2008). Also, several studies have shown that present-day climate does not seem to strongly constrain the future (Stainforth *et al.* 2005; Knutti *et al.* 2006; Sanderson *et al.* 2008). Models continue to improve in simulating what we observe (Reichler & Kim 2008), but do not strongly converge in their projections, indicating that we do not understand what the model metric tells us about the skill of the projections. An illustration of that is given in figure 3, which compares four of the CMIP3 models (*a*–*c*) and the average of all 22 models (*d*–*f*) for simulated absolute present-day boreal winter temperature (*a*,*d*), the bias of it to ERA40 data (*b*,*e*), and the projected warming for the end of the century (*c*,*f*) in the SRES A1B scenario (Nakicenovic & Swart 2000). While the large-scale patterns of absolute temperature are similar, regional biases can be large. However, more importantly, we poorly understand how model bias in the present transfers into differences in projections. The bias of the multi-model mean is typically smaller than the bias of an individual model, but not by as much as one would expect if the models were truly independent (Tebaldi & Knutti 2007; Jun *et al.* 2008*a*). Despite the model improvements in simulating the present-day climatology, the model spread for the projected warming has not substantially decreased in recent years (Knutti *et al.* 2008).

The issues and caveats discussed above should not imply that the current models are 'bad'; rather, they demonstrate that uncertainties are difficult to reduce, and that the definition of model performance is vague. Despite some limitations, climate models have reached a level of maturity that is remarkable. They simulate an ever-increasing range of processes and feedbacks and are tested in a wide range of applications and for different climate states. No credible model has been produced that questions the strong anthropogenic influence on climate in the past and future. I, therefore, argue that the large-scale model projections are very likely robust and accurate within the stated uncertainties. Consistency across models and from one model generation to the next can help to establish which aspects are credible and where it is too early to interpret the results. Recent coordinated experiments and intercomparisons have helped enormously in understanding model differences and quantifying uncertainty. However, model evaluation and the decision of where to allocate resources is still largely done by expert judgement. More formal methods and metrics are needed to quantify progress and uncertainty and to complement expert judgement. Climate models are not even close to providing all the information (e.g. changes in extreme precipitation on local scales) that would be useful, and clearly need to improve to provide projections that are useful for adaptation.

6. So how can we go forward?

One may see the evolution of climate models such as the natural selection of organisms. Successful components or pieces of models are kept and less effective ones are replaced. However, how large and diverse should the zoo of models be? There are at least two competing ways to allocate resources. Model diversity helps to quantify uncertainty, and may increase the chances to discover something new or very different, while steady model improvement of a few existing models helps to improve the fitness for a particular purpose but may be less likely to change things dramatically. With few exceptions (Stainforth *et al.* 2005), AOGCM modelling has traditionally taken the latter approach, in which a single model in each institution is made as complex as can be afforded (Randall *et al.* 2007).

The motivation to use models can be to understand the climate system, to quantify feedbacks, to test hypotheses or to make projections. Model development at least for AOGCMs is partly guided by curiosity and the attempt to learn about the system. The focus on process understanding might partly explain the attempt to reproduce reality with a single model as accurately as possible. Research to improve models will still benefit from that traditional approach in the future. Studying the model response to a perturbation has always been a central part of research, but funding agencies, the IPCC and model intercomparison such as CMIP3 are currently pushing these towards 'near-operational' climate predictions and projections, from seasonal to centennial time scales. The requirements for these in terms of the ideal model (or families of models) may be quite different from those for the 'curiosity-driven' modelling activities.

I argue that more resources are needed to understand model results, and to develop frameworks to quantify uncertainty, model performance, compare different models and communicate results. Many scientists work on model development and assessment, trying to understand the climate (which no doubt

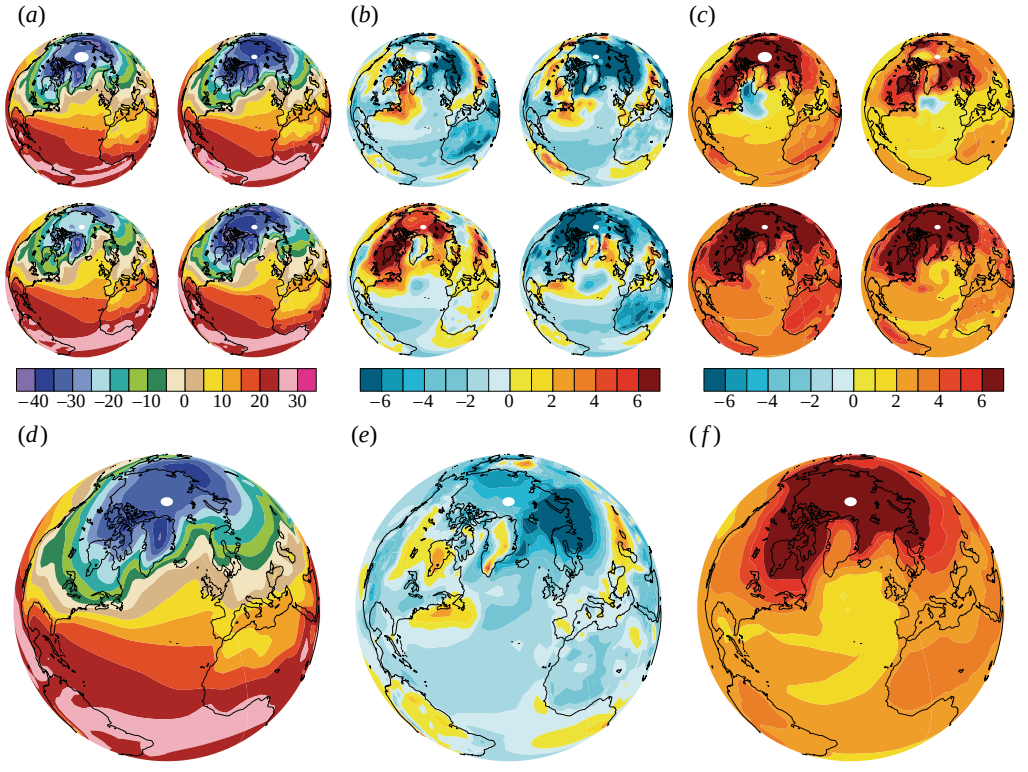

Figure 3. Simulated present-day (1980–1999 average) (a,d) absolute December to February surface temperature, (b,e) the bias of it from ERA 40 reanalysis and (c,f) the projected warming 2080–2099 relative to 1980–1999 for the A1B scenario. Four individual models are shown in $(a–c)$, and the average of 22 CMIP3 models is given in the large panels in $(d–f)$. There is no obvious connection between the present-day bias and the simulated warming.

is crucial) and to find out what the next model should be. Only a handful of people think about how to make the best use of the enormous amounts of data, how to synthesize data for the non-expert, how to effectively communicate the results and how to characterize uncertainty. The amount of data and computing power grows exponentially, but the decision relevance of the model results clearly does not. Huge efforts go into model development, and enormous computer resources are spent to run the models, but fundamental questions on how best to use the results are not answered. For example, the scale at which model projections are robust is poorly known. Methods to quantify uncertainty are still in their infancy, and few methods have been proposed to combine results from multiple models beyond simple averaging. Doing so requires metrics of model performance, a thorny but unavoidable step (Tebaldi & Knutti 2007). These figures of merit or metrics need to relate demonstrably to the projected quantity of interest. A good representation of ENSO is probably irrelevant to predict global temperature a century into the future, so a metric needs to tell us something about the quantity we want to predict. Such metrics can be found, for example, by process understanding or by examining correlations between observables and the projected quantity of interest across a large ensemble of different models (Hall & Qu 2006; Knutti *et al.* 2006). For the latter, the

assumption is of course that the correlation reflects not only the simplicity and uniformity of similar parametrizations built into all models but also some intrinsic behaviour of the natural system, e.g. the magnitude of a process being important on both short (observable) and long (predicted) time scales. Model diversity is critical here in that it helps to understand what makes the projection of two models agree or disagree.

Climate models and their projections may improve in various ways. Biases in atmospheric models appearing days after initialization of a run are often similar to climatological biases. Therefore, the idea of seamless prediction, bridging the gap between short-term forecast (weather and seasonal prediction) and long-term projections, may be promising. The idea is that some processes are relevant for both seasonal and decadal forecasts, and evaluating models on the former may improve the simulation of the latter (Palmer *et al.* 2008). Evaluating models for different climatic states, variability and trends and on abrupt changes observed in the past can reveal model limitations. Initialization with observations (Smith *et al.* 2007), assimilation of data or synchronization of multiple models (or models and data; Kirtman & Shukla 2002) can potentially improve short-term predictions. Better observations, in particular long-term records of observed changes, will further constrain the models and help to understand processes critical to improving model performance.

A hierarchy of different models (Held 2005) and families of similar models can be used to identify which quantities will be most useful to identify model deficiencies and to constrain future projections (e.g. Hall & Qu 2006). Finally, we should be creative in thinking about new types of observations that may be useful (e.g. GPS data; Ringer & Healy 2008).

However, uncertainty in climate projections may still not decrease quickly, both because the present provides only a weak constraint on the future, and because models continue to include more processes and feedbacks interactively, giving rise to new sources of uncertainty and possibilities for model spread. Some uncertainties are even intrinsic and irreducible (volcanic eruptions in the future, the chaotic nature of short-term changes, potentially also some tipping points). Scientists need to specify all possible outcomes, rather than trying to reduce spread where it cannot be reduced. Perturbed physics ensembles with different models are needed to better separate parametric and structural uncertainty. The former can possibly be reduced by calibration; the latter is harder to eliminate. Constraints will emerge as climate change proceeds, so, even without improving our models, we expect our uncertainties to shrink somewhat. In some cases they may also grow, if the additional data reveal that the model is imperfect.

7. Conclusions

Some may argue that long-term forecasts are useless because they cannot be properly evaluated and little can be learnt from a prediction without verification. Indeed, the climate change problem is peculiar in that the past offers no direct well-observed analogy to learn from. Yet I argue that, despite the caveats noted above, climate model projections provide valuable information. The model's ability to reproduce the current climate, the recent observed trends as well as the more distant past, the fact that they are based on physical principles, and the

fact that we can understand and interpret many of the results from known processes provide support for the model's credibility, at least for large scales and certain variables. The value of projections is increased where multiple models are available, in that they indicate which changes are more certain than others. Some scenarios and changes in the climate system are inherently better constrained (e.g. short-term warming trends relate more linearly to observed trends than equilibrium warming), and multiple models and hierarchies of models can help to flag the areas where results depend strongly on model assumptions. Models can also help to infer how much more we may know in a few years by treating one model as reality and predicting it with the other models. Is it more effective to wait for better information or to act earlier? What policy options would we lose by waiting 5 or 10 years? Model development can be guided by understanding why models disagree in their forecasts, what processes and parameters are relevant, and what observations would constrain parameters and therefore constrain projections.

If the goal is to provide decision-relevant information to end-users and policy-makers, then their needs should help to define the focus of model development and climate research. What variables do policy-makers care about? What spatial scale and what time scale are of interest? The decisions of where to allocate computational resources (model resolution, number of model components, model complexity, number of ensemble members, number and type of scenarios and number of models) in CMIP3 were mostly pragmatic, guided by scientific curiosity, experience from the past and computational constraints, rather than by an assessment of what would be most useful for the user and policy-maker (note that it is unclear in this case whether the target user was one studying climate change impacts and adaptation or a climate modeller seeking to improve the model).

On a different level, there is the issue of communicating results. There is a delicate balance between giving the most detailed information possible to guide policy versus communicating only what is known with high confidence. In the former case, all results are used, but there is a risk of the science losing its credibility if the forecasts made a few years later are entirely different or if a forecast made a few years earlier is not verified. The other option is to communicate only what we are confident about. But being conservative (i.e. not being wrong by not saying anything) may be dangerous in this context; once we are sure about certain threats, it may be too late to act. The role of the IPCC scientific assessments has always been to provide the best science to help inform policy decisions, but has become increasingly important in driving policy, for better or worse. Thus the communication of climate projections, their uncertainties and caveats is crucial, and certainly merits more attention.

Finally, as a thought experiment, let us assume that we had a perfect model to make a prediction with no uncertainty. Would the world be any different? Would we more effectively fight the climate change problem? Accurate information on the expected trends is critical for local adaptation, and uncertainties in climate model projections are admittedly an issue. But they are unlikely to be the limiting factor that prevents us from making a decision and acting on, rather than talking about, the climate change problem.

Fruitful discussions with a large number of colleagues have shaped the ideas presented in this paper. In particular, comments from Wendy Parker, Lenny Smith, Bjorn Stevens and two

anonymous reviewers have helped to improve the manuscript. I acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, US Department of Energy.

References

Barnett, T. *et al.* 2005 Detecting and attributing external influences on the climate system: a review of recent advances. *J. Clim.* **18**, 1291–1314. (doi:10.1175/JCLI3329.1)

Bony, S. *et al.* 2006 How well do we understand and evaluate climate change feedback processes? *J. Clim.* **19**, 3445–3482. (doi:10.1175/JCLI3819.1)

Box, G. E. P. 1979 Robustness in the strategy of scientific model building. In *Robustness in statistics* (eds R. L. Launer & G. N. Wilkinson). New York, NY: Academic Press.

Claussen, M. *et al.* 2002 Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Clim. Dyn.* **18**, 579–586. (doi:10.1007/s00382-001-0200-1)

Cox, T. F. & Cox, M. A. A. 2001 *Multidimensional scaling*. London, UK: Chapman & Hall.

Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J., Dix, M., Noda, A., Senior, C. A., Raper, S. & Yap, K. S. 2001 Projections of future climate change. In *Climate change 2001: the scientific basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell & C. A. Johnson), pp. 525–582. Cambridge, UK: Cambridge University Press.

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. (doi:10.1126/science.1064419)

Friedlingstein, P. *et al.* 2006 Climate–carbon cycle feedback analysis: results from the C^4MIP model intercomparison. *J. Clim.* **19**, 3337–3353. (doi:10.1175/JCLI3800.1)

Furrer, R., Knutti, R., Sain, S. R., Nychka, D. W. & Meehl, G. A. 2007 Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.* **34**, L06711. (doi:10.1029/2006GL027754)

Gleckler, P. J., Taylor, K. E. & Doutriaux, C. 2008 Performance metrics for climate models. *J. Geophys. Res. Atmos.* **113**, D06104. (doi:10.1029/2007JD008972)

Hall, A. & Qu, X. 2006 Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.* **33**, L03502. (doi:10.1029/2005GL025127)

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, C., Marengo Orsini, J. A., Nicholls, N., Penner, J. E. & Stott, P. A. 2007 Understanding and attributing climate change. In *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), pp. 663–745. Cambridge, UK: Cambridge University Press.

Held, I. M. 2005 The gap between simulation and understanding in climate modeling. *Bull. Am. Meteorol. Soc.* **80**, 1609–1614. (doi:10.1175/BAMS-86-11-1609)

IPCC 2007 Summary for policymakers. In *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller). Cambridge, UK: Cambridge University Press.

Jansen, E. *et al.* 2007 Paleoclimate. In *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), pp. 433–497. Cambridge, UK: Cambridge University Press.

Jones, P. D., New, M., Parker, D. E., Martin, S. & Rigor, I. G. 1999 Surface air temperature and its variations over the last 150 years. *Rev. Geophys.* **37**, 173–199. (doi:10.1029/1999RG900002)

Jun, M., Knutti, R. & Nychka, D. W. 2008*a* Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *J. Am. Stat. Assoc.* **103**, 934–947. (doi:10.1198/016214507000001265)

Jun, M., Knutti, R. & Nychka, D. W. 2008*b* Local eigenvalue analysis of CMIP3 climate model errors. *Tellus.* **60**, 992–1000. (doi:10.1111/j.1600-0870.2008.00356.x)

Kennedy, M. & O'Hagan, A. 2001 Bayesian calibration of computer models. *J. R. Stat. Soc.* **63B**, 425–464. (doi:10.1111/1467-9868.00294)

Kirtman, B. P. & Shukla, J. 2002 Interactive coupled ensemble: a new coupling strategy for CGCMs. *Geophys. Res. Lett.* **29**, 1367. (doi:10.1029/2002GL014834)

Knutti, R. 2008 Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.* **35**, L18704. (doi:10.1029/2008GL034932)

Knutti, R. & Stocker, T. F. 2002 Limited predictability of the future thermohaline circulation close to an instability threshold. *J. Clim.* **15**, 179–186. (doi:10.1175/1520-0442(2002)015<0179:LPOTFT>2.0.CO;2)

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. (doi:10.1038/416719a)

Knutti, R., Meehl, G. A., 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)

Knutti, R. *et al.* 2008 A review of uncertainties in global temperature projections over the twenty-first century. *J. Clim.* **21**, 2651–2663. (doi:10.1175/2007JCLI2119.1)

Kriegler, E. & Held, H. 2005 Utilizing belief functions for the estimation of future climate change. *Int. J. Approx. Reason.* **39**, 185–209. (doi:10.1016/j.ijar.2004.10.005)

Levitus, S., Antonov, J. I., Boyer, T. P. & Stephens, C. 2000 Warming of the world ocean. *Science* **287**, 2225–2229. (doi:10.1126/science.287.5461.2225)

McWilliams, J. C. 2007 Irreducible imprecision in atmospheric and oceanic simulations. *Proc. Natl Acad. Sci. USA* **104**, 8709–8713. (doi:10.1073/pnas.0702971104)

Medeiros, B., Stevens, B., Held, I. M., Zhao, M., Williamson, D. L., Olson, J. G. & Bretherton, C. S. 2008 Aquaplanets, climate sensitivity, and low clouds. *J. Clim.* **21**, 4974–4991. (doi:10.1175/2008JCLI1995.1)

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J. & Taylor, K. E. 2007*a* The WCRP CMIP3 multimodel dataset—a new era in climate change research. *Bull. Am. Meteorol. Soc.* **88**, 1383–1394. (doi:10.1175/BAMS-88-9-1383)

Meehl, G. A. *et al.* 2007*b* Global climate projections. In *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), pp. 747–845. Cambridge, UK: Cambridge University Press.

Nakicenovic, N. & Swart, R. 2000 *Special report on emissions scenarios.* Cambridge, UK: Cambridge University Press.

Oreskes, N. 2003 The role of quantitative models in science. In *Models in ecosystem science* (eds C. D. Canham, J. J. Cole & W. Lauenroth), pp. 13–31. Princeton, NJ: Princeton University Press.

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. N., Doblas-Reyes, F. J., Weisheimer, A. & Rodwell, M. J. 2008 Toward seamless prediction: calibration of climate change projections using seasonal forecasts. *Bull. Am. Meteorol. Soc.* **89**, 459–470. (doi:10.1175/BAMS-89-4-459)

Parker, W. 2006 Understanding pluralism in climate modeling. *Found. Sci.* **11**, 349–368. (doi:10.1007/s10699-005-3196-x)

Räisänen, J. 2007 How reliable are climate models? *Tellus, Ser. A, Dynam. Meteorol. Oceanogr.* **59**, 2–29. (doi:10.1111/j.1600-0870.2006.00211.x)

Randall, D. A. *et al.* 2007 Climate models and their evaluation. In *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the*

Intergovernmental Panel on Climate Change (eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor & H. L. Miller), pp. 589–662. Cambridge, UK: Cambridge University Press.

Reichler, T. & Kim, J. 2008 How well do coupled models simulate today's climate? *Bull. Am. Meteorol. Soc.* **89**, 303–311. (doi:10.1175/BAMS-89-3-303)

Ringer, M. A. & Healy, S. B. 2008 Monitoring twenty-first century climate using GPS radio occultation bending angles. *Geophys. Res. Lett.* **35**, L05708. (doi:10.1029/2007GL032462)

Sanderson, B. M. *et al.* 2008 Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *J. Clim.* **21**, 2384–2400. (doi:10.1175/2008JCLI1869.1)

Santer, B. D. *et al.* 2003 Influence of satellite data uncertainties on the detection of externally forced climate change. *Science* **300**, 1280–1284. (doi:10.1126/science.1082393)

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R. & Murphy, J. M. 2007 Improved surface temperature prediction for the coming decade from a global climate model. *Science* **317**, 796–799. (doi:10.1126/science.1139540)

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

Stainforth, D. A., Allen, M. R., Tredger, E. R. & Smith, L. A. 2007 Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. A* **365**, 2145–2161. (doi:10.1098/rsta.2007.2074)

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)

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)

Tebaldi, C. & Knutti, R. 2007 The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. Soc. A* **365**, 2053–2075. (doi:10.1098/rsta.2007.2076)

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

Tebaldi, C., Hayhoe, K., Arblaster, J. M. & Meehl, G. A. 2006 Going to the extremes: an intercomparison of model-simulated historical and future changes in extreme events. *Clim. Change* **79**, 185–211. (doi:10.1007/s10584-006-9051-4)

Thompson, D. W. J., Kennedy, J. J., Wallace, J. M. & Jones, P. D. 2008 A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature* **453**, 646–649. (doi:10.1038/nature06982)

van Oldenborgh, G. J., Philip, S. Y. & Collins, M. 2005 El Niño in a changing climate: a multi-model study. *Ocean Sci.* **1**, 81–95.

AUTHOR PROFILE

Reto Knutti

Reto Knutti was born in Saanen, Switzerland, and studied physics at the University of Bern. Fascinated by weather and climate, numerical simulations and high-performance computing, he specialized in the field of climate modelling, in which he earned his PhD from the University of Bern in 2002. Subsequently, he worked as a postdoctoral researcher and visiting scientist at the National Center for Atmospheric Research in Boulder, Colorado, USA, where he still has an affiliation. Since 2007, Reto Knutti has been assistant professor in climate physics at the Institute for Atmospheric and Climate Science at ETH Zurich, Switzerland.

Reto Knutti's research focuses on changes in the global climate system caused by the growing emissions of anthropogenic greenhouse gases such as carbon dioxide. He uses numerical models of different complexity, from simple energy balance to three-dimensional coupled climate models that resolve the atmosphere, ocean, land, sea ice and their interactions. In particular, he develops methods to constrain important feedback processes in the climate system by comparing observations with model results. He uses Bayesian methods to estimate the probability density functions of future global and regional climate change, based on small sets of comprehensive models, very large ensembles of simpler models and statistical emulators such as neural networks. Numerous publications resulting from his work have contributed substantially to a better understanding of the uncertainties in climate projections.

Reto Knutti is a member of the International Detection and Attribution Group and lead author in the latest Fourth Assessment Report of the IPCC. With their advisory activities and their reports, these institutions provide the scientific basis for international agreements on preventing and mitigating climate change.