



Skill and uncertainty in climate models

Julia C. Hargreaves*

Analyses of skill are widely used for assessing weather predictions, but the time scale and lack of validation data mean that it is not generally possible to investigate the predictive skill of today's climate models on the multidecadal time scale. The predictions made with early climate models can, however, be analyzed, and here we show that one such forecast did have skill. It seems reasonable to expect that predictions based on today's more advanced models will be at least as skillful. In general, assessments of predictions based on today's climate models should use Bayesian methods, in which the inevitable subjective decisions are made explicit. For the AR4, the Intergovernmental Panel on Climate Change (IPCC) recommended the Bayesian paradigm for making estimates of uncertainty and probabilistic statements, and here we analyze the way in which uncertainty was actually addressed in the report. Analysis of the ensemble of general circulation models (GCMs) used in the last IPCC report suggests there is little evidence to support the popular notion that the multimodel ensemble is underdispersive, which would imply that the spread of the ensemble may be a reasonable starting point for estimating uncertainty. It is important that the field of uncertainty estimation is developed in order that the best use is made of current scientific knowledge in making predictions of future climate. At the same time, it is only by better understanding the processes and inclusion of these processes in the models, the best estimates of future climate will be closer to the truth. © 2010 John Wiley & Sons, Ltd. *WIREs Clim Change*

In order to make probabilistic predictions of future climate, the quality of forecasts from climate models must be assessed. We have, however, no future climate with which to validate today's models, and the expectation is that future climate is out of sample compared with both the historical period, and those paleoclimate epochs for which we have the most robust proxy records. Consequently, traditional techniques commonly used in numerical weather prediction (NWP), which require multiple realizations of the forecast period, such as calculations of forecast skill, cannot readily be used. Here, we compare the concepts of model skill and uncertainty and calculate the skill for one of the few climate predictions for which this is possible. For predictions being made today, the out of sample nature of future climate change means that there is a sufficiently high degree of

uncertainty in the predictions that Bayesian methods in which the uncertainty can be made explicit, and so tested for robustness, are a necessity. The Intergovernmental Panel on Climate Change (IPCC) reports have had considerable influence on the way uncertainty is considered and communicated. In this paper, we discuss the recommendations made to the authors and the different approaches taken by the chapter authors of the 4th assessment report (AR4). We further consider the potential for using the AR4 general circulation models (GCMs) in probabilistic calculations.

THE DIFFERENCE BETWEEN SKILL AND UNCERTAINTY

Skill

In the NWP context, skill is generally defined as the performance of a particular forecast system in comparison with some other reference technique—usually a simple null hypothesis such as persistence—to see

*Correspondence to: jules@jamstec.go.jp

RIGC/JAMSTEC, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa, 236-0001, Japan

DOI: 10.1002/wcc.58

which has lower errors. A typical example would be the definition from the AMS glossary¹:

Skill: A statistical evaluation of the accuracy of forecasts or the effectiveness of detection techniques. Several simple formulations are commonly used in meteorology. The skill score (SS) is useful for evaluating predictions of temperatures, pressures, or the numerical values of other parameters. It compares a forecaster's root-mean-squared or mean-absolute prediction errors, E_f , over a period of time, with those of a reference technique, E_{refr} , such as forecasts based entirely on climatology or persistence, which involve no analysis of synoptic weather conditions:

$$SS = 1 - \left(\frac{E_f}{E_{refr}} \right) \quad (1)$$

If $SS > 0$, the forecaster or technique is deemed to possess some skill compared with the reference technique.

It is hard to apply this sort of approach to long-term predictions from climate models as there are very few opportunities to compare model predictions with actual outcomes. Although some climate scientists have used the term skill to refer to how well the model field for some climatological variable compares with the null hypothesis of a flat mean field, this is not an equivalent test of skill to that used in NWP as it does not evaluate the model's predictive capability against independent observations. There is, however, one historical example of a prediction that we may analyze along these conventional lines. In 1988, James Hansen presented to the US Congress what was perhaps the first prediction of global climate change over decadal time scales using a numerical climate model, a work which was later published as Hansen et al.² We now present an analysis of this prediction.

The Hansen Forecast

Climate model results in the IPCC reports are generally presented as projections, which are defined as predictions conditional upon the forcing scenario. Several forcing scenarios (including greenhouse gas and aerosol concentrations or emissions) are run for each model, and no indication is given as to which scenario may be most likely. However, of the three model simulations generated (based on three different forcing scenarios) in the work presented by Hansen, one was described as the most realistic, with the others primarily treated as sensitivity tests. This trajectory for the forcings has also turned out rather close to observations. We consider it reasonable, therefore, to conduct a test of the skill treating the scenario and projection from this central case ('Scenario B' which is

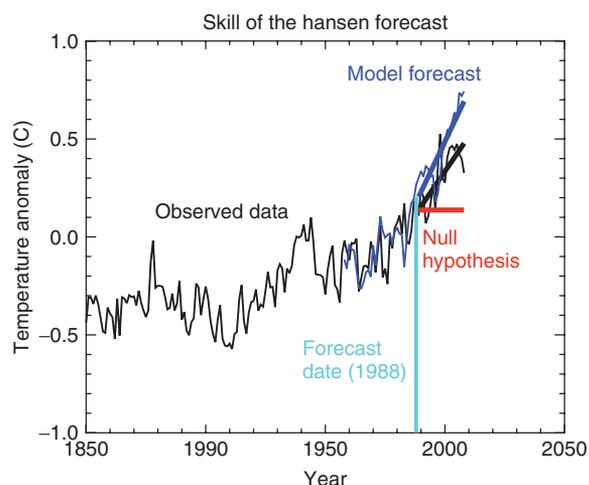


FIGURE 1 | Forecast of Hansen et al.² (blue line) evaluated against observational data (black). Twenty-year trends of forecast and observations are indicated with thick lines, as is the null hypothesis of zero trend (red).

plotted in Figure 1), together as a prediction. Although it would be interesting to consider other variables and a more regional analysis, efforts to reproduce the original model runs have not yet been successful, and the only output which is available from this run is the globally averaged surface temperature anomaly.

There is no question of the model predicting the actual year-to-year temperatures which depend on internal variability rather than external forcing, so we consider only the trend over the 20-year period (1989–2008). Hansen et al.³ and Rahmstorf et al.⁴ have discussed this and similar model runs from the IPCC 1990 report in largely qualitative terms. Here, we perform a simple quantitative evaluation of skill and add some probabilistic evaluation.

The model predicted a trend of $0.26\text{ }^{\circ}\text{C}/\text{decade}$ compared with the observed result of $0.18\text{ }^{\circ}\text{C}/\text{decade}$ (HadCRUT⁵). The choice of baseline for the null hypothesis is not automatic, and there appear to be two natural alternatives: either a persistence forecast, in which future temperatures are predicted to be same as the recent past; or an extrapolation of the recent trend over some historical interval. In both cases, some judgment is required concerning the length of historical interval used. These baseline methods were evaluated over the historical record, by using a segment of the temperature record to establish the baseline for persistence or trend, and then evaluating the performance of this naive forecast against the subsequent 20 years of data. When the two alternatives are evaluated over the historical temperature record in this way, persistence turns out to have the best performance. That is, over the

historical record, it is generally better to use the mean of a 20-year period as a predictor of the subsequent 20 years than to extrapolate a trend forward. This result is robust to different lengths of hindcast and forecast intervals in the range of 5–30 years tested. Therefore, we use persistence (with a 20-year historical baseline) as the null hypothesis. The skill of the model forecast of the global temperature trend according to Eq. 1 is then 0.56, which is substantially greater than zero, indicating that the prediction was skillful.

To evaluate the statistical significance of this result, we can again bootstrap over the historical data to determine the variability of climate trends over 20-year periods. Only 3% of the 120 overlapping 20-year intervals from 1850 to 1988 had a trend as great as that observed, which implies that if the variation of the climate was really caused solely by natural variability with external forcing playing a negligible role, it is very unlikely (probability <5%) that reality would have measured up so well against the forecast. A detailed analysis using detection and attribution techniques would provide a rather more robust and thorough evaluation of this result.⁶

There are several factors that contribute to the Hansen forecast being a little too high. The model had an equilibrium sensitivity to doubled CO₂ of 4.2 °C which is toward the high end of the range considered likely today. Furthermore, the simplified nature of the ocean component of the model results in a transient climate response that is even more extreme. This model also omits the cooling effect of sulfate aerosols, which were poorly understood at that time. Therefore, we can reasonably hope for modern GCMs to give predictions of globally averaged annual temperature that are at least as skillful over similar time spans in the future, although this cannot yet be directly tested against independent data in the manner we have presented here. Furthermore, today's models have more interacting components and are of higher resolution, so the global temperature response is only a first-order test of model performance.

Uncertainty

As there is no direct way of assessing the predictive skill of today's climate models, researchers instead tend to talk in terms of uncertainty in the models and predictions. In contrast with skill, the concept of uncertainty in climate science has neither such an unambiguous definition nor a clear mathematical framework, although an explicitly Bayesian framework is increasingly popular.

Computer models of the climate system are intended to be numerical representations of the

processes which govern climate variability and change. They can be run with different forcing conditions to make predictive statements about parts of the climate system for which there is, as yet, no observational data. This could take the form of variables which have not been measured, paleoclimates and paleoclimate change for which there is not yet much data, and future change for which there is no data at all. They are, however, subject to computational constraints and so are limited in resolution and complexity. Many choices must be made when constructing a climate model about the way processes are represented (due to computational cost, they must be parameterized) and then, once the parameterizations are chosen, choices must also be made for the values of the multitude of parameters, which are often not well known. It is not computationally feasible to test all possible choices, so after constructing a working model utilizing scientific expertise to make the decisions, uncertainty remains as to whether the model is an adequate representation of our understanding of the climate system. That there is still substantial uncertainty can most clearly be seen in the spread of results from the ensembles created through the Model Intercomparison Projects (MIPs). These ensembles are created by each member of the project submitting to the database predefined sets of variables from runs made with their own models using common boundary conditions. There have been over 30 such MIPs over the past couple of decades, focusing on different aspects of the earth system. The spread in the ensembles produced may be considered a guide to the uncertainty in the scientific community's understanding of the behavior of the climate system, and in particular its response to certain forcings. Through analysis of the MIP databases, and comparison with data, the ability of the models to reproduce different aspects of the climate system for present and past climates, and the relative strengths and weaknesses of the different models can be assessed. While large parametric sensitivity experiments with simpler models and intermediate complexity models (EMICs⁷) are relatively plentiful, a smaller number of studies have assessed parametric uncertainty in individual GCMs.^{8–10}

No matter how much the models are improved in their ability to reproduce past and present climates, additional uncertainty arises in the model projections of future climate change due to the possibility that processes important for future climate change have been completely omitted. Some processes are excluded from models, either because they are considered to have a small effect or because scientific understanding is too low for them to be reasonably incorporated into the numerics of the model. Others may be entirely

unknown. While testing the models against data that has not been used in the construction of the model, particularly for climates quite different from today, can build confidence in the predictive ability of the model, the magnitude of this uncertainty cannot readily be defined, and thus remains a subjective component of any probabilistic calculation. It is, therefore, natural to adopt the formalism of Bayesian rather than frequentist statistics, as Bayesian statistics are based on the interpretation of probability as the degree of belief.¹¹ Use of the Bayesian paradigm makes it clear that there exists no such thing as ‘the uncertainty’, alluded to, for example, in Chapter 11 of the AR4.¹² Instead, in Bayesian statistics we have ‘our uncertainty’; there is recognition of the subjective elements, and the final probabilistic statement is a property of the researcher doing the calculation. An advantage of the Bayesian paradigm is that all the uncertainties can be made explicit in the calculation,²⁸ enabling the robustness of the calculation to the assumptions to be calculated.

UNCERTAINTY IN THE AR4

As a basis for an overview of uncertainty in the context of the recent state of the art in climate modeling, we use the IPCC AR4.¹³

Because of the desire to communicate uncertainty in a coherent way to policymakers which is consistent across the three working groups (i.e., across all disciplines related to climate change studies), the IPCC report process has caused some thought to be given to the way uncertainties are expressed. The guidance to authors has evolved somewhat,^{14,15} with the AR4 guidance primarily recommending a Bayesian approach, and that in the case where there is sufficient consensus and evidence for a quantitative assessment, either the terms ‘confidence’ or ‘likelihood’ should be used, where confidence expresses the judgment as to the ‘correctness of a model, an analysis or a statement’, and likelihood relates to a probabilistic assessment about whether something will occur in the future. Although the word ‘correctness’ is used in the IPCC guidance, strictly speaking, statements about how likely a model is to be correct are not meaningful since a model is always an approximation and so is never correct. Perhaps, ‘adequacy’ would have been a better choice of word. This choice of word would also make clearer the subjectivity implicit in the assessment. One other potential confusion of the system proposed by the IPCC is the use of the word ‘likelihood’ to express probability, concepts which are not synonymous in the Bayesian framework. There is, however, some similarity here with the aspects of

uncertainty we identified in the previous section and it would seem that ‘confidence’ in these terms may be more appropriate for expressing the uncertainty as to whether the models adequately represent the climate system as we understand it, while ‘likelihood’ may be used when assessing uncertainty in the climate model projections of future change.

A search of the text shows that Chapter 8 of the AR4 (*Climate models and their evaluation*¹⁶) uses words related to confidence twice as often as those related to likelihood, although it completely avoids the calibrated language recommended by IPCC.¹⁵ This suggests that the authors of that chapter did not find it easy to put a quantitative value on the confidence in the models themselves, preferring instead to talk in general terms of increasing confidence relative to previous generations of models. In contrast, Chapters 10 and 11 of the AR4 (Global and regional model projections^{12,17}) mention words related to likelihood five or six times more often than those related to confidence, and freely use the quantitative language associated with this paradigm. While it is understandable that such a dichotomy may arise when chapters are written independently, this is a somewhat uncomfortable situation since quantification of the confidence in models would appear to be a prerequisite for making probabilistic Bayesian calculations based on the models.

Chapter 8 of the AR4 concentrates principally on the physical climate system, which is mostly built on well-understood physical principles. When we move to look at Chapter 7 (Couplings between changes in the climate system and biogeochemistry¹⁸), things are rather different. Many aspects of biogeochemistry are highly uncertain, and the models rely to a greater extent on empirically derived relationships rather than physical laws. In addition, there is arguably less useful data with which to constrain the models. As a result, the addition of biogeochemistry components to climate models has the effect of increasing the uncertainty in both the models and their predictions. A promising way to approach this problem might be to focus research on the areas of greatest uncertainty while building models in such a way that they can be more readily compared with data that are available.¹⁹

Some consideration has been given as to how well the AR4 models represent uncertainty, and whether the distribution should be wider or narrower. The question we are addressing here is not whether the uncertainty is ‘correct’ in some physical sense, but rather whether the ensemble of projections from the AR4 can be reasonably interpreted as a probabilistic prediction that provides at least an approximate description of our beliefs about the climate system.

One simple method to evaluate ensemble spread can be drawn from the field of NWP. If the ensemble is 'reliable' in the (standard) sense that the truth can be treated as being drawn from the same distribution as the ensemble members, then when n models are ordered according to any statistic of interest, thus dividing the number line up into $n + 1$ intervals (including the two semi-open ends), we should expect observations to fall equiprobably in each interval, this being the basis of the rank histogram evaluation of Talagrand et al.²⁰ For an underdispersive ensemble, the rank histogram of observations will too frequently lie towards the edges or outside the bounds of the ensemble, giving a U-shaped rank histogram, and for an overdispersive ensemble, the rank histogram will have a central dome. Chapter 8 of the AR4¹⁶ contains an overview of the GCM ensemble performance compared with many aspects of the physical climate system, including the mean climate state, temporal variability, extremes, and abrupt change. It seems clear that in most if not all cases, the observational estimates lie toward the median of the range of the ensemble spread, which suggests an overdispersive ensemble. A more rigorous analysis²¹ which analyzed rank histograms derived from the CMIP3 ensemble of temperature, precipitation, and sea level pressure for both annual and seasonal means supports this suggestion. If one makes the (large) assumption that there is a robust relationship between past and future model performance, then the indication is that it may be justifiable to make more precise probabilistic predictions than those made by a direct interpretation of model spread.

INCREASING CONFIDENCE IN CLIMATE FORECASTS

Models can never be perfect,^{22,23} so there will always be uncertainty in their predictions. The production of a climate model projection with no uncertainty estimate is, therefore, of limited value.²⁴ While in principle it would be possible to use ensemble methods to explore the parametric uncertainty in individual GCMs, these models are engineered with high complexity and resolution, up to the limits of available computer power, which makes ensemble experiments impractical. On the other hand, increasing computational power has enabled some ensemble experiments to be performed using slightly dated or lower resolution versions of GCMs.^{8–10} While these experiments can be very useful for exploring the sensitivity of models to parametric uncertainty, there may be limitations to their usefulness for producing estimates of uncertainty in climate projections. For example, Yokohata et al.²⁵ showed that ensembles

from single models with varied parameters may exhibit inconsistent behaviors. Furthermore, tracing the behavior of even slightly dated model versions to the latest versions is not always straightforward (Ref 26, Section 5). While the concept of a hierarchy of models²⁷ is attractive, typically different models with different resolutions and complexities have different strengths and weaknesses and may, therefore, be best used for exploring different aspects of the climate system. A computationally less expensive method for investigating uncertainty in individual models may be to build a simpler model designed specifically to emulate components of the more complex model.^{28–30}

As discussed in the previous section, there is some evidence that the CMIP3 ensemble of models is reliable and can therefore be considered to provide a reasonable estimate of our uncertainty. This has, however, only been tested using steady-state climate statistics, and it is not immediately clear how relevant this is to predictions of future change. In order to increase confidence in the ensemble reliability for future scenarios, scientists have started to look at the recent past and paleoclimate simulations and look for correlations with the simulations of future climate projections (e.g., see Refs 31–33). The hope is that if, for some particular variable in some location, two models are found to be similar for both the past and projected future modeled climate, then the goodness of fit to the observations in that time and place can be used as an guide to the goodness of the model projection. Some have used historical data to try to narrow down the CMIP ensemble by assigning a 'metric' to each model based on their performance with respect to data (e.g., see Refs 31 and 33–35). This is equivalent to treating the CMIP ensemble as a prior estimate, and then re-weighting it according to some cost function. If the CMIP ensemble is found to be overdispersive, then this may be a justifiable line of approach, although the possibility of a too narrow ensemble resulting in an over-confident prediction remains a danger unless the resulting probability distribution function can be shown to be robust. It seems that metrics specific to a particular problem may be more likely to find favor than a single universal metric to be used in all circumstances, partly because the weightings given to the different variables included in the metric are subjective, and different models are better at reproducing different aspects of the climate.

Some recent studies have taken a direct approach and used historical data to re-scale or reject the ensemble members that do not reproduce the feature under study adequately.^{36,37} Such an approach is especially powerful when it uses observations that are unlikely to have been used during the model

development.³⁸ Such experiments are fundamentally equivalent to using some sort of metric to re-weight the ensemble and are also open to exhibiting the same problem of false confidence caused by a too narrow ensemble. It is important, therefore, that a theoretical framework for these sort of calculations is developed so that scientists can develop some confidence in using multimodel ensembles to make predictions for future climate change for different anthropogenic scenarios.

Despite the relative paucity of climate data deeper in the past, assessment of models by simulating paleoclimate epochs can do much to increase confidence in the model projections as it is only in the deeper past that climates very different from today have occurred. Here, correlations between simulated past, present, and future climate states in the PMIP and PMIP2 (Paleoclimate MIPs) model ensembles as well as ensembles made from a single model have been analyzed. For this purpose, the Last Glacial Maximum (LGM) has been most intensively studied because it is a time when the climate state and the CO₂ level were both at quasi-equilibrium levels quite different from that of the modern climate. As yet there is little evidence that the LGM can be used to constrain large-scale variables such as climate sensitivity more tightly than the AR4 models already do,^{39,40} but there is hope that using paleoclimate to better understand and model the relevant physical processes, can be expected to improve the individual model predictions.⁴¹ As more and better data become available (for example, the MARGO sea surface temperature⁴²), a more quantitative approach with respect to paleoclimates should become possible. The PMIP community is also extending its reach to include other climate epochs such as the warm, high CO₂ Pliocene epoch,⁴³ and periods of transient climate change, which should be very valuable. Another way in which paleoclimates can prove very valuable is in those few instances where it is possible to make a prediction of the past climate before data for validation becomes available.^{44,45}

While directly relating the results from simpler models to the more complex GCMs appears to be difficult, EMICS can also be used to investigate uncertainty. These models have varying degrees of complexity, but typically run faster, having lower resolution and simpler representation of physical processes, although they may include more coupled components than contemporary state-of-the-art GCMs. The simplified processes make them more highly parameterized, so that, for example, it may be relatively simple to use them to more thoroughly investigate how robust scientific results are to model differences (e.g., see Refs 46–49). Given automatic

methods based on data assimilation techniques, possibly using such computationally lighter models, it is possible in principle to use cross-validation techniques in which the model is tuned to a subset of the data, with the withheld data being used as validation.⁵⁰ While there are limited opportunities for this on the multidecadal time scale of the historical record for which anthropogenic forcing has been significant, this technique may nevertheless play a part in improving confidence in the models.

CONCLUSION

In the first section, it was argued that it is impossible to assess the skill (in the conventional sense) of current climate forecasts. Analysis of the Hansen forecast of 1988 does, however, give reasons to be hopeful that predictions from current climate models are skillful, at least in terms of the globally averaged temperature trend. Uncertainty in climate modeling and climate model predictions was considered, highlighting the importance of using the Bayesian framework to progress from model confidence to probabilistic predictions. The second section summarized the way uncertainty was treated in the last IPCC report, highlighting the difficulty of quantifying model confidence, but finding evidence to suggest that the ensemble of IPCC models provides a useful basis for a probabilistic calculation. One challenge for those studying uncertainty is the ongoing incorporation of additional poorly understood feedbacks in the models which provide more sources of uncertainty to be investigated. In the third section, we discussed recent work on attempts to improve confidence in the models, by further constraining the multimodel ensemble and investigating links between past and future climate changes. In order to be meaningful, estimates of climate change must include uncertainty estimates. At present, it seems that direct use of the CMIP ensemble may be the best route to follow and research is required to develop methods for understanding the behavior of the ensemble of models in a more coherent way. Bayesian predictions of future change will be obtained by combining all lines of evidence: the multimodel ensembles run for past, present, future and transient experiments; additional expert opinion; data from the present day, historical record, and paleoclimates. Although small steps have been made toward this goal,^{51,52} more serious attempts analyzing a broader range of variables than climate sensitivity should be a high priority.

The aim of climate prediction is not just quantifying, but also reducing, our uncertainty. Uncertainty analysis is a powerful, and under utilized,

tool which can place bounds on the state of current knowledge and point the way for future research, but it is only by better understanding the processes and

inclusion of these processes in the models that the best models can provide predictions that are both more credible and closer to the truth.

ACKNOWLEDGEMENTS

J. D. Annan provided considerable assistance by way of discussion and also helped cut the manuscript down to size. J. D. Annan, G. Schmidt, M. Mann, G. Foster, and J. Nieslsen-Gammon were involved in discussion of the analysis of the Hansen forecast, and G. Schmidt digitized the time series. Thanks to S. L. Weber for helpful comments on the manuscript, and to the editor and two reviewers for their useful suggestions.

REFERENCES

- Glickman TS. *Glossary of Meteorology*. Boston, Mass: American Meteorological Society; 2000.
- Hansen J, Fung I, Lacis A, Rind D, Lebedeff S, Ruedy R, Russell G, P Stone. Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model. *J Geophys Res-Atmos* 1988, 93(D8):9341–9364.
- Hansen J, Sato M, Ruedy R, Lo K, Lea DW, Medina-Elizade M. Global temperature change. *Proc Natl Acad Sci* 2006, 103(39):14288.
- Rahmstorf S, Cazenave A, Church JA, Hansen JE, Keeling RF, Parker DE, Somerville RCJ. Recent climate observations compared to projections. *Science* 2007, 316(5825):709.
- Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J Geophys Res* 2006, 111:D12–D1206.
- Hegerl GC, Zwiers FW, Braconnot P, Gillett NP, Luo Y, Marengo Orsini JA, Nicholls N, Penner JE, Stott PA. 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*. Cambridge: Cambridge University Press 2007.
- Weber SL. The utility of earth-system models of intermediate complexity (EMICS). *WIREs Clim Change* 2009, 1:243–252.
- Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 2004, 430:768–772.
- Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, et al. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* 2005, 433:403–406.
- Annan JD, Hargreaves JC, Ohgaito R, Abe-Ouchi A, Emori S. Efficiently constraining climate sensitivity with paleoclimate simulations. *SOLA* 2005, 1:181–184.
- Bernardo JM, Smith AFM. *Bayesian Theory* Chichester: Wiley; 1994.
- Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, Jones R, Kolli RK, Kwon W-T, Laprise R, et al. Regional 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*. Cambridge: Cambridge University Press; 2007.
- Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press; 2007.
- Moss RH, Schneider SH. *Uncertainties in the IPCC TAR: recommendations to lead authors for more consistent assessment and reporting*. Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC; 2000, 33–51.
- IPCC. Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties. IPCC; 2005.
- Randall DA, Wood RA, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, et al. 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*. Cambridge: Cambridge University Press; 2007.
- Meehl GA, Stocker TF, Collins WD, Friedlingstein P, Gaye AT, Gregory JM, Kitoh A, Knutti R, Murphy JM, Noda A, Raper SCB, Watterson IG, Weaver AJ, Zhao Z-C. Global climate projections. In *Climate Change 2007: The Physical Science Basis. Contribution of the Working Group I to the Fourth Assessment Report*

- of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2007.
18. Denman KL, Brasseur G, Chidthaisong A, Ciais P, Cox PM, Dickinson RE, Hauglustaine D, Heinze C, Holland E, Jacob D, et al. Couplings between changes in the climate system and biogeochemistry. 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*. Cambridge: Cambridge University Press; 2007.
 19. Moorcroft PR. How close are we to a predictive science of the biosphere? *Trends Ecol Evol* 2006, 21(7):400–407.
 20. Talagrand O, Vautard R, Strauss B. *Evaluation of probabilistic prediction systems*. In: Proceedings of ECMWF Workshop on Predictability. ECMWF, Shinfield Park: Reading RG2-9AX; 1997, 1–26.
 21. Annan JD, Hargreaves JC. Reliability of the CMIP3 ensemble. *Geophys Res Lett* 2010, 37(2):L02703.
 22. Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 1994, 263(5147):641.
 23. McWilliams JC. Irreducible imprecision in atmospheric and oceanic simulations. *Proc Natl Acad Sci* 2007, 104(21):8709.
 24. Smith LA. What might we learn from climate forecasts? *Proc Natl Acad Sci* 2002, 99(Suppl. 1):2487.
 25. Yokohata T, Webb MJ, Collins M, Williams KD, Yoshimori M, Hargreaves JC, Annan JD. Structural similarities and differences in climate responses to CO₂ increase between two perturbed physics ensembles by general circulation models. *J Clim* 2009, 23:1392–1410.
 26. Hargreaves J, Annan J. On the importance of paleoclimate modelling for improving predictions of future. *Clim Past* 2009, 5:803–814.
 27. Held IM. The gap between simulation and understanding in climate modeling. *Bull Am Meteor Soc* 2005, 86(11):1609–1614.
 28. Rougier JC. Probabilistic inference for future climate using an ensemble of simulator evaluations. *Clim Change* 2007, doi: 10.1007/s10584-006-9156-9.
 29. Holden PB, Edwards NR, Oliver KIC, Lenton TM, Wilkinson RD. A probabilistic calibration of climate sensitivity and terrestrial carbon change in genie-1. *Clim Dyn* 2009, 1–22. doi: 10.1007/s00382-009-0630-8.
 30. Tachiiri K, Hargreaves JC, Annan JD, Oka A, Abe-Ouchi A, Kawamiya K. Development of a system emulating the global carbon cycle in Earth system models. *Geosci Model Dev Discuss* 2010, 3(1):99–180.
 31. Whetton P, Macadam I, Bathols J, O'Grady J. Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophys Res Lett* 2007, 34(14):L14701. doi: 10.1029/2007GL030025.
 32. Abe M, Shiogama H, Hargreaves JC, Annan JD, Nozawa T, Emori S. Correlation between inter-model similarities in spatial pattern for present and projected future mean climate. *SOLA 2009*, 5:133–136.
 33. Räisänen J, Ruokolainen L, Ylhäisi J. Weighting of model results for improving best estimates of climate change. *Clim Dyn* 2009; 1–16. doi: 10.1007/s00382-009-0659-8.
 34. Tebaldi C, Smith RL, Nychka D, Mearns LO. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *J Clim* 2005, 18(10):1524–1540.
 35. Tebaldi C, Knutti R. The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans Royal Soc A: Math Phys Eng Sci* 2007, 365(1857):2053.
 36. Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* 2000, 407(6804):617–620.
 37. Boé J, Hall A, Qu X. September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience* 2009, 2:341–343.
 38. Lipton P. Testing hypotheses: prediction and prejudice. *Science* 2005, 307:219–221.
 39. Crucifix M. Does the last glacial maximum constrain climate sensitivity? *Geophys Res Lett* 2006, 33:L18701. doi: 10.1029/2006GL027137.
 40. Otto-Bliesner BL, Schneider R, Brady EC, Kucera M, Abe-Ouchi A, Bard E, Braconnot P, Crucifix M, Hewitt CD, Kageyama M, et al. A comparison of PMIP2 model simulations and the MARGO proxy reconstruction for tropical sea surface temperatures at last glacial maximum. *Clim Dyn* 2009, 32:799–815. doi: 10.1007/s00382-008-0509-0.
 41. Hargreaves JC, Abe-Ouchi A, Annan JD. Linking glacial and future climates through an ensemble of GCM simulations. *Clim Past* 2007, 3(1):77–87.
 42. Waelbroeck C, Paul A, Kucera M, Rosell-Melé A, Weinelt M, Schneider R, Mix AC, Abelmann A, L Armand Bard E, et al. Constraints on the magnitude and patterns of ocean cooling at the last glacial maximum. *Nature Geosci* 2009, 2:127–132.
 43. Haywood AM, Dowsett HJ, Otto-Bliesner B, Chandler MA, Dolan AM, Hill DJ, Lunt DJ, Robinson MM, Rosenbloom N, Slazmann U, et al. Pliocene model inter-comparison project (PlioMIP): experimental design and boundary conditions (experiment 1). *Geosci Model Dev Discuss* 2010, 3:227–242.
 44. Saltzman B, Verbitsky M. Predicting the Vostok CO₂ curve. *Nature* 1995, 377:690.

45. Wolff EW, Chappellaz J, Fischer H, Kull C, Miller H, Stocker TF, Watson AJ. The epic challenge to the earth system modeling community. *EOS Trans Am Geophys Union* 2004, 85(35):363.
46. Forest CE, Stone PH, Sokolov AP, Allen MR, Webster MD. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* 2002, 295(5552):113–117.
47. Knutti R, Stocker TF, Joos F, Plattner G-K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* 2002, 416:719–723.
48. Hargreaves JC, Annan JD, Edwards NR, Marsh R. An efficient climate forecasting method using an intermediate complexity Earth system model and the ensemble Kalman filter. *Clim Dyn* 2004, 23(7–8):745–760.
49. Ridgwell A, Hargreaves JC, Edwards NR, Annan JD, Lenton TM, Marsh R, Yool A, Watson A. Marine geochemical data assimilation in an efficient earth system model of global biogeochemical cycling. *Biogeosciences* 2007, 4(1):87–104.
50. Hargreaves JC, Annan JD. Assimilation of paleo-data in a simple Earth system model. *Clim Dyn* 2002, 19(5–6): 371–381.
51. Annan JD, Hargreaves JC. Using multiple observationally-based constraints to estimate climate sensitivity. *Geophys Res Lett* 2006, 33(L06704). doi: 10.1029/2005GL025259.
52. Hegerl GC, Crowley TJ, Hyde WT, Frame DJ. Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature* 2006, 440:1029–1032.