Climatic Change (2010) 102:395–404
DOI 10.1007/s10584-010-9800-2

EDITORIAL

# The end of model democracy?

## An editorial comment

Reto Knutti

Received: 24 November 2009 / Accepted: 4 January 2010 / Published online: 13 January 2010

## 1 The trillion dollar garden party—an analogy

Imagine you are hosting a garden party tomorrow and you are trying to decide whether or not to put up a tent against the rain. You read the weather forecast in the newspaper and you ask the farmer next door, and you look at the sky (knowing that persistence is often not a bad weather forecast). So you get three predictions, but how would you aggregate them? Would you average them with equal weight? You might trust the forecast model more (or less) than the farmer, not because you understand how either of them generates their prediction, but because of your past experience in similar situations. But why seek advice from more than one source in the first place? We intuitively assume that the combined information from multiple sources improves our understanding and therefore our ability to decide. Now having read one newspaper forecast already, would a second and a third one increase your confidence? That seems unlikely, because you know that all newspaper forecasts are based on one of only a few numerical weather prediction models. Now once you have decided on a set of forecasts, and irrespective of whether they agree or not, you will have to synthesize the different pieces of information and decide about the tent for the party. The optimal decision probably involves more than just the most likely prediction. If the damage without the tent is likely to be large, and if putting up the tent is easy, then you might go for the tent in a case of large prediction uncertainty even if the most likely outcome is no rain.

Although it may seem far-fetched at first, the problem of climate projection is in fact similar in many respects to the garden party situation discussed above. So far, projections from multiple climate models were often aggregated into simple averages, standard deviations and ranges. One example is the recent Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC),

R. Knutti (✉)
Institute for Atmospheric and Climate Science, ETH Zurich,
Universitatstrasse 16, 8092 Zurich, Switzerland
e-mail: reto.knutti@env.ethz.ch

which was based largely on multi-model averages of the models participating in the World Climate Research Project (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) (Meehl et al. 2007). But is this the best use of the information? Or are we finally ready to move beyond the 'one-model-one-vote' approach? Some recent publications suggest that 'the end of model democracy' (a quote first used by Vladimir Kattsov at an IPCC meeting in 2006) may be near, but the problem is far from trivial. In this issue, Smith and Chandler (2010) propose that for rainfall over the Murray Darling basin in south east Australia, present-day precipitation mean and variability are useful indicators for the evaluation of models, and they find that models performing well today show a more similar trend in the future. At least in their case, eliminating poor models therefore decreases the spread of the ensemble. Although this is not always the case, similar findings for Alaska and Greenland are reported by Walsh et al. (2008). Ideas of down-weighting or eliminating models have been around for a while (Giorgi and Mearns 2002, 2003) but the widespread availability of perturbed physics and multi-model ensembles has sparked more interest in the community for methods to evaluate, combine and possibly weight models, and constrain projections using observations (Annan et al. 2005; Eyring et al. 2007; Forest et al. 2002; Furrer et al. 2007a, b; Greene et al. 2006; Hall and Qu 2006; Hargreaves et al. 2004; Jun et al. 2008a, b; Knutti 2008a; Knutti et al. 2009; Lopez et al. 2006; Murphy et al. 2004; Piani et al. 2005; Räisänen 2005, 2007; Sanderson et al. 2008; Schmittner et al. 2005; Shukla et al. 2006; Tebaldi and Knutti 2007; Tebaldi et al. 2004, 2005; Tebaldi and Sanso 2009; van Oldenborgh et al. 2005; Weigel et al. 2008). In the following section, I try to summarize some of the pertinent questions that in my view remain unresolved in combining multiple models.

There are of course several aspects where the analogy with the garden party above will fail. First, we are working with a forecast system that strictly speaking has never been proven to have skill, nor to be wrong, at least on the time scales of centuries. Second, we are dealing with a rather expensive garden party that involves billions of people and trillions of dollars. The decisions based on our forecasts might shape the world of the future, so a prediction that is overconfident might be an expensive failure. And finally, it is not just about deciding yes or no as in the case of the tent, but about deciding on one of many possible strategies based on an admittedly incomplete understanding of an extremely complex system. Accordingly, providing a clear recommendation for a way forward is far from trivial.

## 2 Making sense of multiple models

In some applications, the use of multiple models has been shown to provide predictions superior to those of a single model (Cantelaube and Terres 2005; Doblas-Reyes et al. 2003; Hagedorn et al. 2005; Palmer et al. 2005; Thomson et al. 2006). Surprisingly, even adding a poor model can improve a prediction if the individual models tend to be overconfident (Weigel et al. 2008). So at first, it seems obvious that we will benefit if we can use the good aspects of all models and eliminate the bad parts. In addition, multiple predictions would give us some information about robustness or uncertainty in the prediction. Unfortunately, things are harder than they seem at first sight.

### 2.1 Ensemble design and spread

The first issue to understand is the design of the model ensemble. A scientific experiment usually requires control over the setup. Such control is possible in so-called perturbed physics ensembles (Murphy et al. 2004; Stainforth et al. 2005) where a particular model is run multiple times with different parameter settings to sample the parametric uncertainties. But all results in such perturbed physics ensembles (PPE) depend on the structure of the underlying model. If that model for example does not contain a carbon cycle, then none of the perturbed ensemble members will sample that uncertainty. The alternative is so-called multi-model ensembles, sets of different models collected in coordinated model intercomparisons like CMIP3. The problem in the case of CMIP3 is that the sample of models is neither random nor systematic. Since anyone can contribute, it is simply unclear how to interpret the set of models in the first place. But because most modeling groups only contribute their “best” model, all of them carefully calibrated to the same observations, and no group is deliberately trying to push their model to extreme behavior, it is possible that the multi-model ensemble underestimates the uncertainty in the climate system. Although this is hard to confirm or reject, there may even be an element of ‘social anchoring’ and a tendency towards consensus: for quantities that cannot be measured (for example climate sensitivity), it is easier to be in the middle of the crowd than far outside.

### 2.2 Model independence

The CMIP3 dataset contains simulations from about 25 general circulation models, which seems quite a large sample given the cost of developing and running these types of models. But how independent are these models really? As discussed in the introduction, a second or third weather forecast for the garden party is redundant if it comes from the same model or the same expert. A second opinion (assuming it agrees with the first) should only boost our confidence if it is based on at least partly independent information. Characterizing the dependence of climate model projections is difficult, simply because the models cannot be evaluated directly on the projections. But recent work suggests that the behavior of the CMIP3 multi-model ensemble may be similar to a set of probably only five to ten independent models (Jun et al. 2008a, b; Knutti et al. 2009). This doesn’t come as a surprise, as several institutions have contributed a set of two or three models sharing parts of the code, input datasets and certainly expertise of those developing the model. Because parts of the model biases are similar in some or all models, including many more models of the same quality should not make us more confident about their projections. Many statistical methods however do assume independence of the models (e.g. Furrer et al. 2007b; Tebaldi et al. 2005), which results in a reduction in uncertainty with an increasing number of models (Knutti et al. 2009; Tebaldi and Knutti 2007) unlikely to be meaningful at least for a large number of models. An alternative to the idea of each model approximating the true climate with some error is to consider each model trajectory as a possible trajectory for the Earth (Tebaldi and Sanso 2009). In this case the multi model mean does not converge to the truth.

### 2.3 'All models are wrong', or the issue of structural uncertainty

The fact that no parameter set in a model matches all observations means that the model is structurally wrong, i.e. there are processes that are not adequately resolved or parameterized, or missing entirely (Kennedy and O'Hagan 2001; McWilliams 2007; Smith 2002). Incorporating structural uncertainty (or model discrepancy) in an uncertainty analysis is a tough problem. Some argue that structural problems are too big compared to the observational uncertainty, implying that all models are so wrong that we cannot even attach likelihoods to models (Stainforth et al. 2007). I would not go that far. All models can be shown to be inaccurate to some degree if we use enough data to evaluate them. But this may not matter in some cases, and is expected because a model is only an approximate description of the real system. We construct airplanes with computer models without being able to properly simulate turbulence, yet the airplanes fly as expected. So a model serves its purpose if it makes a useful and reliable prediction, even if its structure is simple. In fact the beauty of a model often is its simplicity, and the fact that we can understand it and relate it to the behavior of more complex models (Held 2005). Structural problems that are similar across many models however place a limit on the confidence we obtain from robustness. Some model results are perfectly robust... yet wrong.

### 2.4 How many models, and how to combine them?

Given the uncertainties in observations and limitations in our understanding of the climate system, there are many ways to build a model that can be justified. Therefore, different models are generally interpreted as complementary rather than incompatible (Parker 2006). But how many models do we need? State of the art climate models are expensive to run. Therefore, a compromise between the number of model versions, the number of simulations, and the complexity of the models is needed. Those solely interested in the most likely outcome may argue for only a few very complex and complete models, while those trying to quantify uncertainty would want a large number of models pushing the edges of what is plausible. More models lead to larger spread unless data can constrain the likely range and eliminate outliers. Too few models on the other hand can quickly lead to overconfidence if they are based on similar assumptions. Finally, averaging models leads to unwanted effects like smoothing of spatially heterogeneous patterns, so it is unclear whether an average across models is physically meaningful at all (Knutti et al. 2009).

### 2.5 Model evaluation and metrics

One of the main difficulties is to define the criteria to separate a 'good' and a 'bad' model to establish the credibility of a projection (Knutti 2008a). For a weather forecast, we know the next day whether we were right or wrong, so skill can be quantified by repeated verification of forecasts. For climate, the lifetime of the model and the time scale on which decisions need to be made (typically a few years) are much shorter than the lead time of the forecast (decades to centuries), and direct verification of the forecast is impossible. So we have to establish credibility of the models by evaluating them on present-day climatology, variability, past anthropogenic trends, paleoclimate or by evaluating processes (Bony et al. 2006; Knutti

2008a; Räisänen 2007; Randall et al. 2007). In some cases, past trends are strongly related to future trends, e.g. for large-scale greenhouse-gas-induced warming (Stott and Kettleborough 2002) or Arctic sea ice decline (Boe et al. 2009). The study by Smith and Chandler (2010) in this issue shows that present-day climate and variability are related to the predicted change in precipitation in parts of Australia. But in most cases such relations between obvious metrics of observable quantities and projections are in fact weak or non-existent (Knutti et al. 2009). There is an infinite number of ways to define a metric, and which one to use is likely to depend on the application. A model with a realistic simulation of the Indian Monsoon for example is not necessarily a good model to predict Arctic sea ice decline.

### 2.6 Model calibration and evaluation

Model development and model evaluation partly use the same datasets, which raises the question of circular reasoning. Indeed climate models are getting better in reproducing what we observe, e.g. the mean state of climate (Gleckler et al. 2008; Räisänen 2007; Randall et al. 2007; Reichler and Kim 2008), or past trends (Hegerl et al. 2007); but the spread in projections is not obviously decreasing (Knutti et al. 2008). This could partly be the result of including more processes, which introduce more degrees of freedom for the projection, but it could also reflect what is known as model calibration or tuning. The views of whether or how much to tune a model differ greatly, but it is clear that independent datasets to evaluate models are rare, and sometimes one can get the right result for the wrong reason. While models are clearly getting more comprehensive, it is unclear how much of the convergence with observations is due to better understanding of the processes, and increased realism and agreement in how they are modeled, and how much is tuning. Such tuning to observations can also lead to apparent correlation without obvious plausible causes. For example, models match the observed warming with different combinations of aerosol forcing and climate sensitivities (Kiehl 2007; Knutti 2008b), so the two appear to be negatively correlated even though there is no obvious causal relationship.

## 3 Conclusions

Some studies have recently proposed to down-weight or eliminate some ‘bad’ climate models, recalibrate projections or estimate uncertainties based on metrics of skill (e.g. Boe et al. 2009; Eyring et al. 2005; Giorgi and Mearns 2003; Perkins and Pitman 2009; Perkins et al. 2009; Schmittner et al. 2005; Smith et al. 2009; Tebaldi et al. 2005; Whetton et al. 2007). A few have found only small differences when measures of model skill are included (Murphy et al. 2004; Santer et al. 2009; Schmittner et al. 2005). In other cases the relations of observables and projections were strong (Boe et al. 2009; Hall and Qu 2006) and supported by physical arguments, but many standard quantities (e.g. climatological surface temperature and precipitation) do not clearly constrain the CMIP3 models and weighting is not straightforward (Knutti et al. 2009).

In summary, I believe that the problem of weighting and evaluating models proves to be harder than expected. Here I offer some personal recommendations in how we might improve on the current situation:

- The community would benefit from a larger set of proposed methods and metrics, including process evaluation, multivariate methods, new statistical methods (e.g. machine learning) as the basis for the assessment in the Fifth Assessment Report of IPCC. We learn from the diversity of models, and we will learn from different ways to evaluate and combine them. With the next model intercomparisons around the corner (CMIP5), a new playground will soon be available to test ideas.
- New methods are needed to cope with a more diverse collection of models (e.g. some with and some without carbon cycle or chemistry) and to combine multimodel ensembles with (potentially large) perturbed physics ensembles (Murphy et al. 2007).
- Metrics and criteria for model evaluation must be demonstrated to relate to the projection. While an overall metric of skill can be defined (Gleckler et al. 2008; Reichler and Kim 2008), specific projections (e.g. El Niño) will likely require individual assessments.
- It may be less controversial to downweight or eliminate a few very poor models that are clearly unable to mimic important processes, or are even physically implausible (e.g. violation of conservation of momentum or mass), than to agree on the best model. By only focusing on the latest generation of models we are in fact eliminating older model versions, often without justification.
- The interplay of model calibration and evaluation must be clarified. Agreement with observations is often (and maybe misleadingly) used to demonstrate progress (Randall et al. 2007; Reichler and Kim 2008) even if it might partly result from tuning or compensating errors.
- Structural uncertainty in climate models is large for some variables but often ignored. Improved methods are required for the design of coordinated experiments and for the generation of probabilistic projections that take into account structural uncertainties.
- Uncertainties will be underestimated if the ensemble is too narrow to begin with. This would explain the lack of clear observational constraints: observations used in model development have little additional value to weight or select models. We would probably learn more from an extreme model and its potential failure than from yet another shot in the center of the consensus range. A credible uncertainty estimate requires the extremes to be sampled and tested.
- Model evaluation on common metrics is useful but diversity of metrics is important. Process understanding must complement 'broad brush metrics' that simply sum up squared errors. Consensus on metrics carries the danger of a 'dog and pony show' which encourages convergence even if there is no reason for.
- Not sampling or down-weighting extreme behavior and rewarding convergence will lead to overconfidence and will reward models that are similar, even if the similarity is just a result of similar errors, assumptions or model structure. The benefit of a more narrow projection must be compared to the potential damage of overconfident projections and wrong adaptation decisions resulting from it.
- Model evaluation, detection and attribution, observations and projections are intimately linked. Assessments like that by the IPCC would benefit from a tighter

integration of these topics. This is far from simple but seems key to a better quantification of uncertainty in projections.

- Our ability to produce data grows faster than the technology to store and transfer it (or even analyze and understand it). Analysis of the expected one Petabyte of data from CMIP5 (due in 2010/2011) will be near impossible without novel approaches like server-side processing and effective data reduction before transmission. But even with those technologies, data analysis will be a time consuming, tedious and expensive task that could be severely limited by time and technical resources, in particular for those not located at wealthy institutions.

The ultimate goal of our efforts is of course to quantify and reduce uncertainties. But defining progress only by our ability to narrow model spread is dangerous in my view. It prevents a comprehensive uncertainty analysis and encourages scientists to make oversimplified assumptions, to use excessive model weighting, to neglect sources of uncertainty and to a priori discount model behavior that does not support the consensus. There is a real danger of convergence as a result of tuning, consensus on metrics and peer pressure, rather than improved understanding and models. Climate projections are inherently uncertain, and part of the uncertainty related to variability is irreducible (Hawking and Sutton 2009). Uncertainty quantification is critical, and the approach of running the most expensive model once on the largest computer available may not be very helpful for that. A recent modeling summit even argued for an integration of everything from weather to seasonal to century timescale predictions as well as data assimilation into a huge very high resolution model and asked for three orders of magnitude more computing power (http://wcrp.wmo.int/documents/WCRP_WorldModellingSummit_Jan2009.pdf). I acknowledge that some of these ideas are interesting, but remain skeptical about the overall benefit of the 'one size fits all' model. Similarly, while I agree that the progress in technology is crucial, a factor of 1,000 in computing power alone (which will be achieved in a mere 11 years even if past trends just persist) will not help much unless we can significantly improve process understanding and models.

The challenge of extracting the most useful information and estimating uncertainties from the large number of models and simulations is still ahead of us. I predict that despite the fact that climate models contain more realistic representations of more processes, uncertainties in climate projections will not decrease substantially in the near future, but could in fact increase due to added complexity. This emphasizes the need for decisions that are robust against alternative future climate outcomes (Dessai et al. 2009; Lempert and Schlesinger 2000) and that allow adjustments when new evidence emerges. But as with the weather forecast for the garden party, my prediction is uncertain. Different members of the 'multi scientist ensemble' may have alternative views.

## References

Annan JD, Hargreaves JC, Edwards NR, Marsh R (2005) Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. Ocean Model 8:135–154

Boe JL, Hall A, Qu X (2009) September sea-ice cover in the Arctic Ocean projected to vanish by 2100. Nature Geosci 2:341–343

Bony S, Colman R, Kattsov VM, Allan RP, Bretherton CS, Dufresne J-L, Hall A, Hallegatte S, Holland MM, Ingram W, Randall DA, Soden BJ, Tselioudis G, Webb MJ (2006) How well do we understand and evaluate climate change feedback processes? J Clim 19:3445–3482

Cantelaube P, Terres JM (2005) Seasonal weather forecasts for crop yield modelling in Europe. Tellus A 57:476–487

Dessai S, Hulme M, Lempert R, Pielke RA Jr (2009) Do we need better predictions to adapt to a changing climate? EOS 90:111–112

Doblas-Reyes FJ, Pavan V, Stephenson DB (2003) The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. Clim Dyn 21:501–514

Eyring V, Harris NRP, Rex M, Shepherd TG, Fahey DW, Amanatidis GT, Austin J, Chipperfield MP, Dameris M, Forster PMF, Gettelman A, Graf HF, Nagashima T, Newman PA, Pawson S, Prather MJ, Pyle JA, Salawitch RJ, Santer BD, Waugh DW (2005) A strategy for process-oriented validation of coupled chemistry-climate models. Bull Am Meteorol Soc 86:1117–1133

Eyring V, Waugh DW, Bodeker GE, Cordero E, Akiyoshi H, Austin J, Beagley SR, Boville BA, Braesicke P, Bruhl C, Butchart N, Chipperfield MP, Dameris M, Deckert R, Deushi M, Frith SM, Garcia RR, Gettelman A, Giorgetta MA, Kinnison DE, Mancini E, Manzini E, Marsh DR, Matthes S, Nagashima T, Newman PA, Nielsen JE, Pawson S, Pitari G, Plummer DA, Rozanov E, Schraner M, Scinocca JF, Semeniuk K, Shepherd TG, Shibata K, Steil B, Stolarski RS, Tian W, Yoshiki M (2007) Multimodel projections of stratospheric ozone in the 21st century. J Geophys Res-Atmos 112:D16303. doi:16310.11029/12006JD008332

Forest CE, Stone PH, Sokolov AP, Allen MR, Webster MD (2002) Quantifying uncertainties in climate system properties with the use of recent climate observations. Science 295:113–117

Furrer R, Sain SR, Nychka D, Meehl GA (2007a) Multivariate Bayesian analysis of atmosphere-ocean general circulation models. Environ Ecol Stat 14:249–266

Furrer R, Knutti R, Sain SR, Nychka DW, Meehl GA (2007b) Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. Geophys Res Lett 34:L06711 doi:10.1029/2006GL027754

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method. J Clim 15:1141–1158

Giorgi F, Mearns LO (2003) Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. Geophys Res Lett 30:1629. doi:1610.1029/2003GL017130

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res-Atmos 113:D06104. doi:10.1029/2007JD008972

Greene AM, Goddard L, Lall U (2006) Probabilistic multimodel regional temperature change projections. J Clim 19:4326–4346

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: basic concept. Tellus 57A:219–233

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:03510.01029/02005GL025127

Hargreaves JC, Annan JD, Edwards NR, Marsh R (2004) An efficient climate forecasting method using an intermediate complexity Earth System Model and the ensemble Kalman filter. Clim Dyn 23:745–760

Hawking E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. Bull Am Meteorol Soc 90:1095–1107. doi:1010.1175/2009BAMS2607.1091

Hegerl GC, Zwiers FW, Braconnot P, Gillett NP, Luo C, Marengo Orsini JA, Nicholls N, Penner JE, Stott PA (2007) Understanding and attributing climate change. In: Solomon S, Quin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) 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 University Press, Cambridge, pp 663–745

Held IM (2005) The gap between simulation and understanding in climate modeling. Bull Am Meteorol Soc 80:1609–1614. doi:1610.1175/BAMS-1686-1611-1609

Jun M, Knutti R, Nychka DW (2008a) Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? J Am Stat Assoc 103:934–947

Jun MY, Knutti R, Nychka DW (2008b) Local eigenvalue analysis of CMIP3 climate model errors. Tellus Ser A Dyn Meteorol Oceanogr 60:992–1000

Kennedy M, O’Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc 63B:425–464

Kiehl JT (2007) Twentieth century climate model response and climate sensitivity. Geophys Res Lett 34:L22710. doi:22710.21029/22007GL031383

Knutti R (2008a) Should we believe model predictions of future climate change? Philos T R Soc A 366:4647–4664

Knutti R (2008b) Why are climate models reproducing the observed global surface warming so well? Geophys Res Lett 35:L18704. doi:18710.11029/12008GL034932

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2009) Challenges in combining projections from multiple models. J Clim (in press)

Knutti R, Allen MR, Friedlingstein P, Gregory JM, Hegerl GC, Meehl GA, Meinshausen M, Murphy JM, Plattner GK, Raper SCB, Stocker TF, Stott PA, Teng H, Wigley TML (2008) A review of uncertainties in global temperature projections over the twenty-first century. J Clim 21:2651–2663

Lempert RJ, Schlesinger ME (2000) Robust strategies for abating climate change—an editorial essay. Clim Change 45:387–401

Lopez A, Tebaldi C, New M, Stainforth DA, Allen MR, Kettleborough JA (2006) Two approaches to quantifying uncertainty in global temperature changes. J Clim 19:4785

McWilliams JC (2007) Irreducible imprecision in atmospheric and oceanic simulations. Proc Natl Acad Sci U S A 104:8709–8713

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multimodel dataset—a new era in climate change research. Bull Am Meteorol Soc 88:1383–1394

Murphy JM, Booth BBB, Collins M, Harris GR, Sexton DMH, Webb MJ (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos Trans R Soc A 365:1993–2028

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 429:768–772

Palmer TN, Doblas-Reyes FJ, Hagedorn R, Weisheimer A (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. Philos Trans R Soc B 360:1991–1998

Parker W (2006) Understanding model pluralism in climate science. Found Sci 11:349–368

Perkins SE, Pitman AJ (2009) Do weak AR4 models bias projections of future climate changes over Australia? Clim Change 93:527–558

Perkins SE, Pitman AJ, Sisson SA (2009) Smaller projected increases in 20-year temperature returns over Australia in skill-selected climate models. Geophys Res Lett 36:L06710

Piani C, Frame DJ, Stainforth DA, Allen MR (2005) Constraints on climate change from a multi-thousand member ensemble of simulations. Geophys Res Lett 32:L23825

Räisänen J (2005) Probability distributions of $CO_2$-induced global warming as inferred directly from multimodel ensemble simulations. Geophysica 41:19–30

Räisänen J (2007) How reliable are climate models? Tellus Ser A Dyn Meteorol Oceanogr 59:2–29

Randall DA, Wood RA, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, Stouffer RJ, Sumi A, Taylor K (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) 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 University Press, Cambridge, pp 589–662

Reichler T, Kim J (2008) How well do coupled models simulate today's climate? Bull Am Meteorol Soc 89:303–311

Sanderson BM, Knutti R, Aina T, Christensen C, Faull N, Frame DJ, Ingram WJ, Piani C, Stainforth DA, Stone DA, Allen MR (2008) Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. J Clim 21:2384–2400

Santer BD, Taylor KE, Gleckler PJ, Bonfils C, Barnett TP, Pierce DW, Wigley TML, Mears C, Wentz FJ, Bruggemann W, Gillett NP, Klein SA, Solomon S, Stott PA, Wehner MF (2009) Incorporating model quality information in climate change detection and attribution studies. Proc Natl Acad Sci U S A 106:14778–14783

Schmittner A, Latif M, Schneider B (2005) Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations. Geophys Res Lett 32:L23710

Shukla J, DelSole T, Fennessy M, Kinter J, Paolino D (2006) Climate model fidelity and projections of climate change. Geophys Res Lett 33:L07702

Smith I, Chandler E (2010) Refining rainfall projections for the Murray Darling Basin of south-east Australia—the effect of sampling model results based on performance. Clim Change (in press)

Smith LA (2002) What might we learn from climate forecasts? Proc Natl Acad Sci U S A 99:2487–2492

Smith RL, Tebaldi C, Nychka DW, Mearns LO (2009) Bayesian modeling of uncertainty in ensembles of climate models. J Am Stat Assoc Appl Case Stud 104:97–116

Stainforth DA, Allen MR, Tredger ER, Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc A 365:2145–2161

Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, Piani C, Sexton D, Smith LA, Spicer RA, Thorpe AJ, Allen MR (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433:403–406

Stott PA, Kettleborough JA (2002) Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. Nature 416:723–726

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 365:2053–2075

Tebaldi C, Sanso B (2009) Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. J R Stat Soc Ser A Stat Soc 172:83–106

Tebaldi C, Mearns LO, Nychka D, Smith RL (2004) Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. Geophys Res Lett 31:L24213

Tebaldi C, Smith RW, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. J Clim 18:1524–1540

Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, Phindela T, Morse AP, Palmer TN (2006) Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439:576–579

van Oldenborgh GJ, Philip SY, Collins M (2005) El Niño in a changing climate: a multi-model study. Oc Sci 1:81–95

Walsh JE, Chapman WL, Romanovsky V, Christensen JH, Stendel M (2008) Global climate model performance over Alaska and Greenland. J Clim 21:6156–6174

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260

Whetton P, Macadam I, Bathols J, O'Grady J (2007) Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. Geophys Res Lett 34:L14701. doi:14710.11029/12007GL030025