Test of a decadal climate forecast

To the Editor — Early climate forecasts[1] are often claimed to have overestimated recent warming. However, their evaluation is challenging for two reasons. First, only a small number of independent forecasts have been made. And second, an independent test of a forecast of the decadal response to external climate forcing requires observations taken over at least one and a half decades from the last observations used to make the forecast, because internally generated climate fluctuations can persist for several years. Here we assess one of the first probabilistic climate forecasts with a full uncertainty assessment[2] that was based on climate models and data up to 1996. Using observations of global temperature over the ensuing 16 years, we find that the original forecast is performing significantly better than a hypothetical alternative based on the assumption that decade-to-decade temperature fluctuations consist of a random walk, that is, a sequence of random fluctuations with no externally driven warming trend. The original climate forecast[2] also outperforms a very simple interpretation of the climate models used for the latest Assessment of the Intergovernmental Panel on Climate Change (IPCC), supporting the conclusions of previous assessments that the spread of such an ensemble is not, on its own, an adequate measure of forecast uncertainty[3].

An evaluation[4] of early predictions of the IPCC noted that although these predictions provide support for the contention that climate is responding to enhanced greenhouse gas levels in accordance with historical expectations, formal evaluation is difficult because these early forecasts were framed as responses to idealized, CO_2-only scenarios and were not couched in unambiguous probabilistic terms.

A climate forecast can only be evaluated and potentially falsified if it provides a quantitative range of uncertainty[5]. For example, if, at verification time, observations lie outside the 5–95% forecast uncertainty range, a forecast can be said to have been falsified at the 10% level. This could indicate an error in initial conditions, forcing or response, or it could occur simply by chance. Hence forecasters must be clear what it is they are forecasting (including uncertainties), and at the same time, evaluators must focus on what has actually been forecast[6]. For example, the disagreement (if any) between recent model simulations and observed climate evolution within the period 1998–2012 would be more significant if the scientific community had previously claimed that these models provided a complete forecast of uncertainty in the distribution of trends over this period[7] — which it did not.

One of the first climate forecasts to provide a formal estimate of the range of uncertainty was a prediction of global mean surface temperature[2] made in 1999 using simulations with the HadCM2

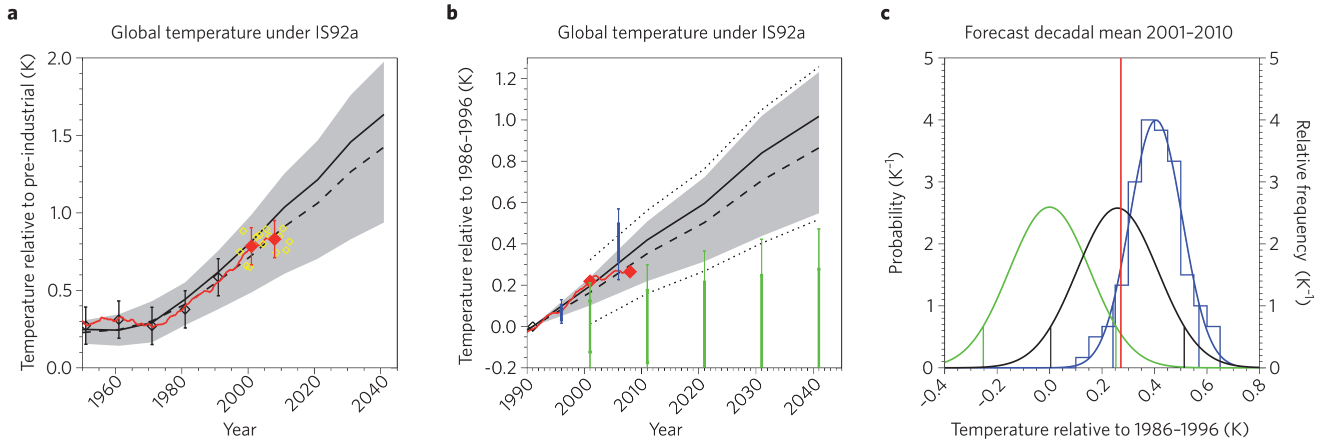

Figure 1 | Evaluation of decadal climate forecasts (updated from Fig. 3 of ref. 2). **a**, Global temperatures relative to the pre-industrial era under a version of the IS92a scenario of relatively high greenhouse gas and anthropogenic sulphate forcing[10]. The solid line shows the original ensemble mean. The grey shaded region indicates the 5–95% uncertainty interval in forecast anthropogenic warming after scaling the model-simulated spatiotemporal patterns of response to greenhouse gas and sulphate forcing to give the best fit (dashed line) to observations over 1946–1996. Large diamonds are decadal means of the observations: open black, used in calibration; solid red, first- and last-available out-of-sample forecast decades. Vertical bars on the black and red diamonds show 5–95% ranges on decadal mean temperatures to be expected from internal variability as simulated by the HadCM2 model, which is also used for uncertainty estimates[14]. It is consistent with more recent models and with residuals of the fit[15]. The red line is a running decadal mean through the updated observations. Yellow diamonds are annual temperatures for the forecast period. **b**, Forecast warming re-expressed relative to 1986–1996, using the original software and constraints[2]. Dotted lines indicate further uncertainty in decadal mean temperatures arising from internal variability estimated from the HadCM2 control, added in quadrature. Observed temperatures are shown relative to 1986–1996, omitting the 18 months following the Pinatubo eruption to avoid confusing anthropogenic warming with the recovery from that eruption. Thin and thick bars show 5–95% and 17–83% forecast ranges, respectively, from an unforced random walk model (green) calibrated with decade-to-decade temperature differences over the observed record until 1996 and from interpreting 120 CMIP5 'historical' simulations as a simple, un-weighted ensemble, also omitting 1992 because of Pinatubo (blue). **c**, Probabilistic forecasts for the decade 2001–2010, relative to 1986–1996, from ref. 2 (black), CMIP5 (blue) and random walk (green). All forecasts represented as Gaussians, with the CMIP5 ensemble histogram also shown. The red vertical line shows verification and the dotted vertical lines show the 5th and 95th forecast percentiles.

climate model (Fig. 1a, solid curve)[8]. The radiative forcing in these simulations represented the effect of all greenhouse gases by an increase of 1% per year in atmospheric CO_2 concentrations from 1990. Together with the impact of sulphate aerosols (which was also included), the total anthropogenic forcing used in the model reached 2 Wm^{-2} in 2010, consistent with current estimates of the actual forcing. The original projection used an ensemble of four simulations with different initial conditions, and no attempt was made to initialize the atmosphere or ocean to conditions of the 1990s.

The best-fit projection of temperature change was obtained by scaling model-simulated climate patterns to observations made from 1946–1996 (Fig. 1a, dashed line). This approach assumes that fractional errors persist, or that a model that has over-predicted observed warming by 20% in the past will continue to do so[2,9]. The downward adjustment in Fig. 1 from black solid to dashed lines indicates that the original simulation was over-predicting the observed response by a small (insignificant) margin. This forecast provided support for the 2001 IPCC report[10], which stated that "anthropogenic warming is likely to lie in the range 0.1–0.2 °C per decade over the next few decades" for this scenario of radiative forcing. Because this forecast was made using data collected until only August 1996 (and submitted in 1999), we can now run an exemplary evaluation using entirely independent data.

The decadal running mean of observed global average temperatures[11] (Fig. 1a, red line) shows a small upward excursion owing to warm years around 2000, followed by a reversion to the original forecast. Observed decadal mean temperatures still lie comfortably within the 5–95% confidence interval when the original 1999 forecast is re-expressed relative to the most recent (and warmest) decade, which was used to constrain the original projection[2], September 1986 to August 1996 (Fig. 1b). Temperatures for individual years lie outside the uncertainty limits, but this does not falsify the prediction of decadal-mean warming.

Rather than evaluating a climate forecast on its own, we can assess whether the observations would be deemed systematically more probable under rival forecasts of varying degrees of sophistication. Fig. 1b shows hypothetical forecasts based on an unforced random walk (green bars) and a very simple interpretation of the 'CMIP5 ensemble' of simulations used in the latest IPCC Assessment[12] (blue bars). In both cases, the verification seems to lie close to the limits of the 5–95% uncertainty ranges.

A more detailed, probabilistic evaluation of these three forecasts is shown in Fig. 1c, focussing on the decadal mean temperature for 2001–2010, relative to the period 1986–1996 (observed value shown by red vertical line). The original 1999 forecast[2] performs best, but the CMIP5 forecast also clearly outperforms the random walk, primarily because it has better sharpness[13]; that is, the forecast distribution is narrower and therefore exhibits a higher probability density at any given percentile. Specifically, the verification (red line) lies just inside the 5–95% range for the CMIP5 forecast, and only just outside this range for the random walk, but the probability density of the CMIP5 forecast at this point is twice as high. This finding illustrates the importance of both calibration and sharpness in forecast evaluation[6]. A very vague forecast is trivially hard to falsify, but will be outperformed by a sharper (more specific) forecast even if they both miss the verification by a similar number of standard errors.

Even if temperatures for the decade 2007–2016 remain no higher than those for the decade 2002–2011, the 1999 forecast[2] would still not be falsified at the 10% confidence level. However, it would no longer be substantially better than the random walk. If, however, temperatures have still not risen above those of the most recent decade by 2017–2026, in the absence of an explosive volcanic eruption, asteroid strike, nuclear exchange or other neglected short-term climate forcing, then the observations will fall outside the range of the dotted lines in Fig. 1b and the forecast[2] will have been falsified at the 10% level.

Apparent falsification of a climate forecast might be caused by temporary errors in either the forcing or response. For example, some of the relatively rapid warming from the 1990s to 2000s in the CMIP5 ensemble of simulations may have been caused by a one-off overestimate in the rate of reduction of cooling sulphur pollution over this period. This overestimate may not persist in the future. Hence, a forecast that has been falsified at some level need not continue to be so. Likewise, failure to falsify a forecast does not guarantee continued success, particularly as the balance of external forcings changes[2]. Nevertheless, many of the sources of error in a climate forecast, such as an over- or underestimate of the sensitivity of the climate to external forcing, may be expected to persist over time. Hence formal out-of-sample evaluation provides a valuable test of our understanding of climate change and, in this instance, illustrates the benefits of forecasts combining modelling with observations.

References

1. Hansen, J. *et al. J. Geophy. Res. Atmos.* **93,** 9341–9364 (1988).
2. Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. *Nature* **407,** 617–620 (2000).
3. Knutti, R, Furrer, R., Tebaldi, C., Cermak, J. & Meehl, G. A. *J. Clim.* **23,** 2739–2758 (2010).
4. Frame, D. J. & Stone, D. A. *Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1763 (2012).
5. Pielke, R. A. *Nature Geosci.* **1,** 206 (2008).
6. Brocker, J. & Smith, L. A. *Weather Forecast.* **22,** 382–388 (2007).
7. Easterling, D. R. & Wehner, M. F. *Geophys Res Lett.* **36,** L08706 (2009).
8. Mitchell, J. F. B., Johns, T. C., Gregory, J. M. & Tett, S. F. B. *Nature* **376,** 501–504 (1995).
9. Lee, T. C. K., Zwiers, F. W., Zhang, X. & Tsao, M. *J. Clim.* **19,** 5305–5318 (2006).
10. IPCC *Climate Change 2001 : The Scientific Basis* (eds Houghton, J. T. *et al.*) (Cambridge Univ. Press, 2001).
11. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. *J. Geophys Res. Atmos.* **117,** D08101 (2012).
12. Taylor, K. E., Stouffer, R. J., Meehl, G. A. *Bull. Am. Meteorol. Soc.* **93,** 485–498 (2012).
13. Gneiting, T., Balabdaoui, F. & Raftery, A. E. *J. R. Stat. Soc. B* **69,** 243–268 (2007).
14. Allen, M. R. & Stott, P. A. *Clim. Dynam.* **21,** 477–491 (2003).
15. Allen, M. R. & Tett, S. F. B. *Clim. Dynam.* **15,** 419–434 (1999).

Acknowledgements

This work was supported by the Joint DECC and Defra Met Office Hadley Centre Climate Programme (GA01101). P.A.S. and M.R.A. were also supported by the US Department of Energy through the International Detection and Attribution Group (IDAG) and M.R.A. received support from the Natural Environment Research Council (NERC) End-to-end Quantification of Uncertainty for Impacts Prediction (EQUIP) project. We thank the climate modelling groups contributing to the World Climate Research Programme (WCRP) CMIP5 project for producing and making available their model output, and G. Jones for providing data.

Myles R. Allen[1]*, John F.B. Mitchell[2] and Peter A. Stott[2]
[1]School of Geography and the Environment, and Department of Physics, University of Oxford, OX1 3QY, Oxford, UK, [2]Met Office, Fitzroy Road, EX1 3PB, Exeter, UK.
*e-mail: myles.allen@ouce.ox.ac.uk