Predictions of 21st century warming constrained by recent climate observations

Daniel J. Rowlands[1,2], David J. Frame[1,2,3], Duncan Ackerley[4], Tolu Aina[5], Ben B. B. Booth[6], Carl Christensen[1], Matthew Collins[6,7], Nicholas Faull[1], Chris E. Forest[8], Benjamin S. Grandey[1], Edward Gryspeerdt[1], Eleanor J. Highwood[9], William J. Ingram[1,6], Sylvia Knight[10], Ana Lopez[11], Neil Massey[1,3], Nicolai Meinshausen[12], Claudio Piani[13,14], Suzanne M. Rosier[1], Benjamin M. Sanderson[15], Leonard A. Smith[11], Dàithi A. Stone[16], Milo Thurston[5], Kuniko Yamazaki[1], Y. Hiro Yamazaki[1,17], and Myles R. Allen[1,2]

[1]Atmospheric, Oceanic & Planetary Physics, Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK.

[2]Oxford University Centre for the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK.

[3]Smith School of Enterprise and the Environment, Hayes House, 75 George St, Oxford OX1 2BQ, UK.

[4]Monash Weather and Climate, Monash University, Clayton, Victoria, 3800, Australia.

[5]Oxford e-Research Centre, Keble Road, Oxford OX1 3QG, UK.

[6]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PU, UK.

[7]College of Engineering, Mathematic and Physical Sciences, University of Exeter, Exeter, EX4 4QJ, UK.

[8]Department of Meteorology, Pennsylvania State University, University Park, PA 16802, USA.

[9]Department of Meteorology, University of Reading, Earley Gate, Reading, RG6 6BB, UK.

[10]Royal Meteorological Society, Reading, RG1 7LL, UK.

[11]Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK.

[12]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.

[13]Abdus Salam International Center for Theoretical Physics, Trieste, Italy.

[14]The American University of Paris, Paris, France.

[15]National Center for Atmospheric Research, 1850 Table Mesa Dr, Boulder, CO 80305, USA.

[16]Climate Systems Analysis Group, University of Cape Town, South Africa.

[17]School of Geography, Politics and Sociology, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.

Forecasts of 21^{st} century climate require physically-based simulation models constrained to be consistent with recent observations, including a systematic estimate of uncertainty. To date, these have relied on scaling approaches[1,2], large ensembles of low dimensional climate models[3,4], or small ensembles of complex coupled atmosphere-ocean general circulation models[5,6] (AOGCMs). Ensembles of opportunity, such as the Coupled Model Inter-comparison Project Phase 3 (CMIP-3)[5], under-represent known uncertainties in key climate system properties derived from independent sources[7–9]. Here we present results from the first multi-thousand member perturbed physics ensemble of transient AOGCM simulations from the climate*prediction*.net BBC climate change experiment (BBC CCE). Model versions consistent with the observed temperature changes over the past 50 years and current uncertainties in global mean top of atmosphere (TOA) flux imbalances show global-mean warming relative to 1961-1990 ranging from 1.4-3K by 2050 (1.9-4.7K by 2075) under a mid-range forcing scenario. This is consistent with results from simpler models and the expert assessment provided in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4)[10], but extends towards larger warming than the models typically used for impact assessments in the CMIP-3 AOGCM ensemble. We therefore provide the first direct AOGCM evidence for high response worlds consistent with recent observed climate change and a mid-range "no mitigation" forcing scenario, with potentially wide ranging implications for the development of robust adaptation policies.

Uncertainties in the global mean temperature response to sustained anthropogenic greenhouse gas forcing are controlled by physical processes responsible for 3 key properties: (1) the equilibrium climate sensitivity, (2) the rate of ocean heat uptake and (3) the historical aerosol forcing[3,4]. In the latest generation of AOGCMs contributing to IPCC AR4, the known uncertainties in these quantities may not have been fully sampled, partially due to a correlation between climate sensitivity and aerosol forcing[7,8], a tendency to overestimate ocean heat uptake[11] and compensation between short-wave and long-wave feedbacks[9]. This complicates the interpretation of the ensemble spread (approximately +/-25%) as a direct uncertainty estimate, a point reflected in the fact that the "likely" (> 66% probability) uncertainty range on the transient response in IPCC AR4 was explicitly, and subjectively, given as -40% to +60% of the CMIP-3 ensemble mean for global mean temperature in 2100. The IPCC expert range was supported by a range of sources[10], including studies using pattern scaling[1,2], ensembles of intermediate-complexity models[3,4] and estimates of the strength of carbon-cycle feedbacks[12]. Thus while the CMIP-3 ensemble is a valuable expression of plausible, physically coherent responses over the coming decades exploring model structural uncertainties, it fails to reflect the full range of uncertainties indicated by expert opinion and other methods.

In the absence of uncertainty guidance or indicators at regional scales, studies have relied on the CMIP-3 ensemble spread as a proxy for response uncertainty[13], or statistical post-processing to correct and inflate uncertainty estimates[14], though this raises the risk of violating the physical constraints provided by dynamical AOGCM simulations, especially when extrapolating beyond the range of behaviour in the raw ensemble.

Perturbed-physics ensembles offer a systematic approach to quantify uncertainty in the climate system response to external forcing. Previous studies have focussed on the equilibrium response[15,16], or have explored uncertainties single components of the climate system such as the atmosphere or ocean[6] under transient forcing. Here we investigate uncertainties in the 21st transient response in a multi-thousand-member ensemble of transient AOGCM simulations from the climate*prediction*.net BBC climate change experiment (BBC CCE). We use HadCM3L, an AOGCM version of the UK Met Office Unified Model, generating ensemble members by perturbing the physics in the atmosphere, ocean and sulphur cycle components (Methods), and applying flux adjustments to correct any imbalances that occur when model atmospheres and oceans are coupled[17].

For each model version two sets of 160 year simulations were performed: (1) control simulations with constant forcing (representative of 1880-1920 mean conditions) to check and allow for unforced drifts and (2) transient simulations from 1920-2080 forced with changes in greenhouse gases and sulphate emissions under the SRES A1B emissions scenario[18], and set of solar and volcanic forcing scenarios (Methods and

Fig. SI 1).

Fig. 1 shows the evolution of global-mean surface temperatures in the BBC CCE (relative to 1961-1990), each coloured by the goodness-of-fit to observations of recent surface temperature changes, as detailed below. The raw ensemble range (1.1-4.1K around 2050) is potentially misleading, since many ensemble members have an unrealistic response to the forcing over the past 50 years. We therefore compare model-simulated spatio-temporal patterns of 5 year mean surface (1.5m) temperatures over 1961-2010 with observations[19], all expressed as anomalies from the respective 1961-1990 mean. We test model versions against temperature changes over the past 50 years because they have been shown to correlate well with future warming[1], whilst mean temperatures do not[20]. We filter the ensemble to retain only model versions requiring a global annual mean flux adjustment in the range $\pm 5W/m^2$, comparable with estimates of observational uncertainty[6], to include a measure of the quality of the model base climatology.

Assessing goodness-of-fit requires a measure of the expected error between model and observations due to sampling uncertainty, primarily from internally-generated climate variability. We estimate this using segments of long pre-industrial control simulations from CMIP-3, filtered to retain spatial scales on which AOGCM-based estimates of variability are reliable (Fig. SI 6).

Weighting model versions explicitly can make results that very sensitive to noise in individual simulations[21] and to parameter sampling design[22]. Although parameter ranges used were informed by expert opinion[15], sampling within these ranges is problematic since many parameters do not have direct real world counterparts. We focus instead on the range of projections provided by model versions that satisfy a given goodness-of-fit threshold: this will be insensitive to sampling design provided the ensemble sufficiently large.

Without a goodness-of-fit (or model error) threshold, hindcasts of 2001-2010 global-mean warming relative to 1961-1990 show a wide range from 0-1.5K (Fig. 2a). We define a 'likely' range (66% confidence interval) by considering the range from ensemble members with model error (y-axis) lower than the 66^{th} percentile of the distribution of model error arising from internal variability alone, estimated from CMIP-3 control segments (black crosses), giving a range of 0.3-0.9K. This is the range of warming to date (relative to 1961-1990) that we estimate might have occurred at this confidence level given the evidence of our ensemble and estimates of internal climate variability from CMIP-3. The observed warming (0.5K – thick black line and grey vertical bar) is close to our best-fit model version (not identical, since we use more than just global mean trend information in our measure of model error), and 0.1K below the centre of our uncertainty range, consistent with temperatures over 2001-2010 being slightly depressed by a combination of internal variability[23] or recent stratospheric water vapour trends[24] and exceptionally low solar minimum[25], neither of which is represented in our ensemble. Note that the grey bar represents uncertainty in the warming that actually occurred, while our constrained ensemble range represents the warming that might have occurred over this period given internal variability and response uncertainty.

On the assumption that models that simulate past changes realistically are our best candidates for making estimates of the future, we can use the same approach to estimate uncertainties in the future climate response. Ensemble members consistent with the observations show a range of warming of 1.4-3K around 2050 under the SRES A1B scenario (Fig. 2b), representing a 66% (or 'likely') confidence interval (Methods).

No ensemble members warm by less than 1K by 2050 under this scenario, despite the large size of the ensemble and allowance for natural forcing uncertainty: we allow explicitly for future volcanic activity and include a scenario in which solar activity falls back to 1900 levels. This is consistent with energy balance considerations[26] given the level of greenhouse gas forcing by 2050 and the lower limit of climate sensitivity explored in the ensemble at approximately 2K, consistent with the lower end of the range of sensitivities considered likely by the IPCC AR4[10].

The lower end of our 66% confidence interval for 2050 warming at 1.4K is consistent with the lowest responses in the CMIP-3 ensemble (filled circles Fig. 2b), lower than the lowest realistic (on this measure) members of the QUMP HadCM3 perturbed physics ensemble[6] (open circles Fig. 2b), and higher than IPCC expert lower bound[10] (the CMIP-3 ensemble-mean minus 40%). This is contingent evidence that the real-world response is likely to be at least as large as the lowest responses in the CMIP-3 ensemble, and that the IPCC AR4 estimate of the lower bound was probably over-conservative. This comparison assumes a constant fractional uncertainty in the 21^{st} century response[1,8], since the IPCC expert estimate was given only for 2100.

At about 3K, the upper end of our uncertainty range for 2050 warming is consistent with both the highest responses in the QUMP ensemble and the IPCC upper estimate of the CMIP-3 ensemble-mean plus 60%[10], but substantially higher than highest responses of the CMIP-3 ensemble members that are generally used for impact assessment (one model gave a higher response, but was not highlighted in headline uncertainty ranges because of concerns about its stability). Thus uncertainty estimates based solely on ensembles-of-opportunity or small perturbed-physics ensembles are likely to be underestimated compared to independent studies[1,4]. We are reluctant to quote a more precise upper bound because of the small number of model versions in this region and the fact that goodness-of-fit does not deteriorate as rapidly as it does at the lower bound, possibly because of the inclusion of natural forcing uncertainty: we can, however, conclude that warming substantially greater than 3K by 2050 is unlikely unless forcing is substantially higher than the A1B scenario[27]. Towards the end of the century, we observe a similar relationship with the IPCC expert estimate (red bar, Fig. 1), although by that time it is likely that the uncertainty would be larger if carbon-cycle feedbacks were included in the BBC CCE[12].

To the extent that policy makers require "a range of plausible representations of future climate"[28] providing uncertainty guidance in this way can have an important role to play. Additional observational constraints may reduce uncertainty further[29], although the application of climatological constraints here is complicated by the use of flux adjustments and the pre-selection of atmospheric configurations with reasonable base climatology. We find little sensitivity in our results to varying the flux adjustment threshold and removing this constraint entirely adds approximately 0.5K to the upper bound in 2050 through admitting a number of high climate sensitivity model versions (Fig. SI 9). Conversely, we are likely to have undersampled uncertainty in ocean heat uptake through perturbing only a single, coarse-resolution, ocean model structure:[6] more generally, sampling structural uncertainty might compensate for the impact of further observational constraints.

Unlike uncertainty estimates based on intermediate-complexity models[11], pattern-scaling[2] or statistical emulation[30], every member of the BBC CCE is consistent with the physical constraints of a 3-D AOGCM, ensuring physical coherence of results for investigating joint uncertainties. Fig. 3 shows surface warming in a low response (Model A, global $\Delta T_{2050} = 1.4K$) and high response (Model B, global $\Delta T_{2050} = 3K$) ensemble member. For 2001-2010, both the observations (Fig. 3a) and models show broadly similar features of enhanced warming over land, which is amplified by 2041-2060. There is a large diversity of regional responses within the sub-ensemble consistent with observations. For example, the range of Pacific equatorial warming (specifically the Niño 3.4 region) relative to warming over the Pacific as a whole between Model A and B is larger than the corresponding range observed in either the CMIP-3 or QUMP ensembles, providing evidence that perturbed-physics ensembles can sample spatial response uncertainty.

Uncertainty estimates for the transient response are conditioned on a given emissions scenario[10]. For the SRES A1B scenario, we have shown that a thorough sampling of uncertainty in key climate system properties and forcings produces a wider range of projections for the coming century consistent with recent observations than in the CMIP-3 ensemble used for regional projections in IPCC AR4, and similar to the IPCC authors' expert assessment of uncertainty in the global response. Reliance on the spread of responses in an ensemble of opportunity can underestimate uncertainties, particularly at the upper end of the range for 21st century warming. The BBC CCE provides a set of physically coherent, physically plausible worlds, beyond the range generated by ensembles of opportunity, which can aid the development of robust climate adaptation policies.

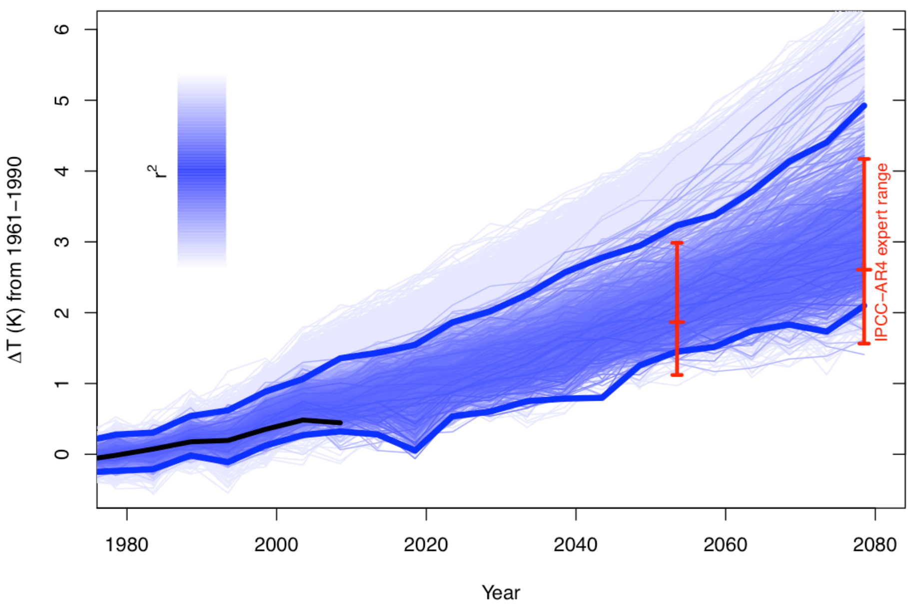

Figure 1: Evolution of uncertainties in global-mean temperature projections under SRES A1B in the BBC CCE. Blue colouring indicates goodness-of-fit between observations and ensemble members, plotted in order of increasing agreement (light to dark blue). Black line, the evolution of observations, and thick blue lines the 'likely' range (66% confidence interval) from the BBC CCE (See text for details). Red bars show the IPCC-AR4 expert 'likely' range around 2050 and 2080. All temperatures are relative to the corresponding 1961-1990 mean.

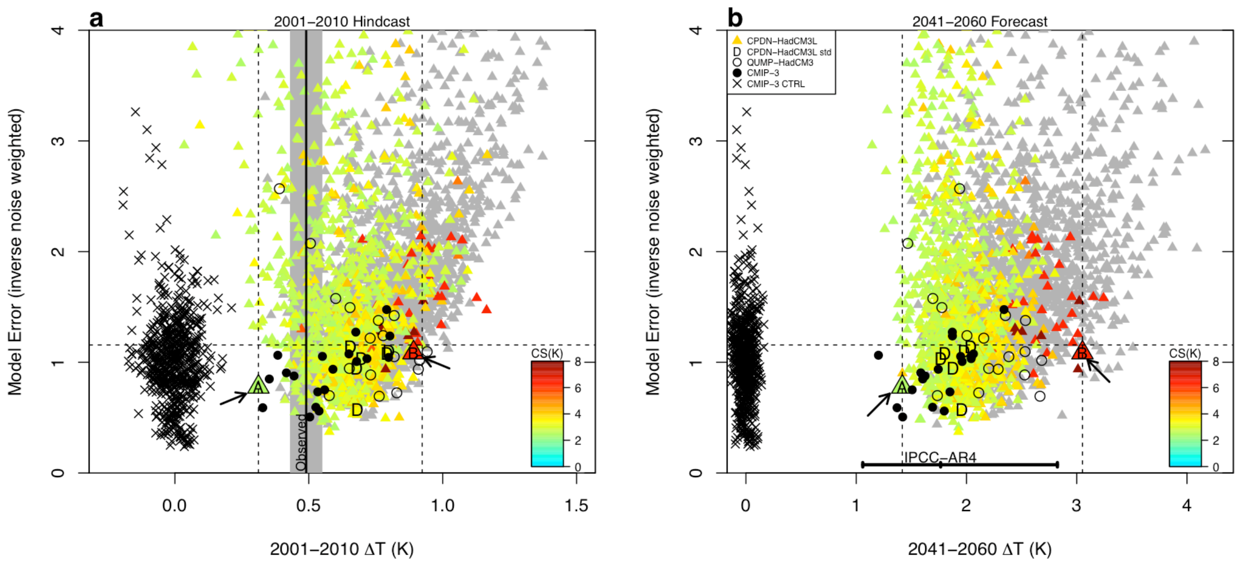

Figure 2: Goodness-of-fit to recent temperature changes as a function of global-mean warming. **a**, 2001-2010 hindcast; **b**, 2041-2060 forecast under SRES A1B for global-mean temperature both as anomalies from 1961-1990. Coloured points, members of the BBC CCE perturbed physics ensemble, with colours denoting the corresponding slab model estimated equilibrium climate sensitivity. Black crosses, realisations of model error and corresponding temperature changes arising from estimates of internal variability for the same periods, with the horizontal line denoting the 66^{th} percentile of the error distribution. Vertical dotted lines, the range of the BBC CCE ensemble with errors lower than this percentile corresponding to a 'likely' range (66% confidence interval). Grey triangles, simulations with global annual mean flux adjustments outside $\pm 5W/m^2$. Black vertical bar and grey band in **a**, observations and 'likely' range. Horizontal bar in **b**, the expert IPCC AR4 'likely' range. Black filled circles CMIP-3 simulations, black open circles QUMP HadCM3 simulations. Arrows and larger triangles refer to models highlighted in Fig. 3.

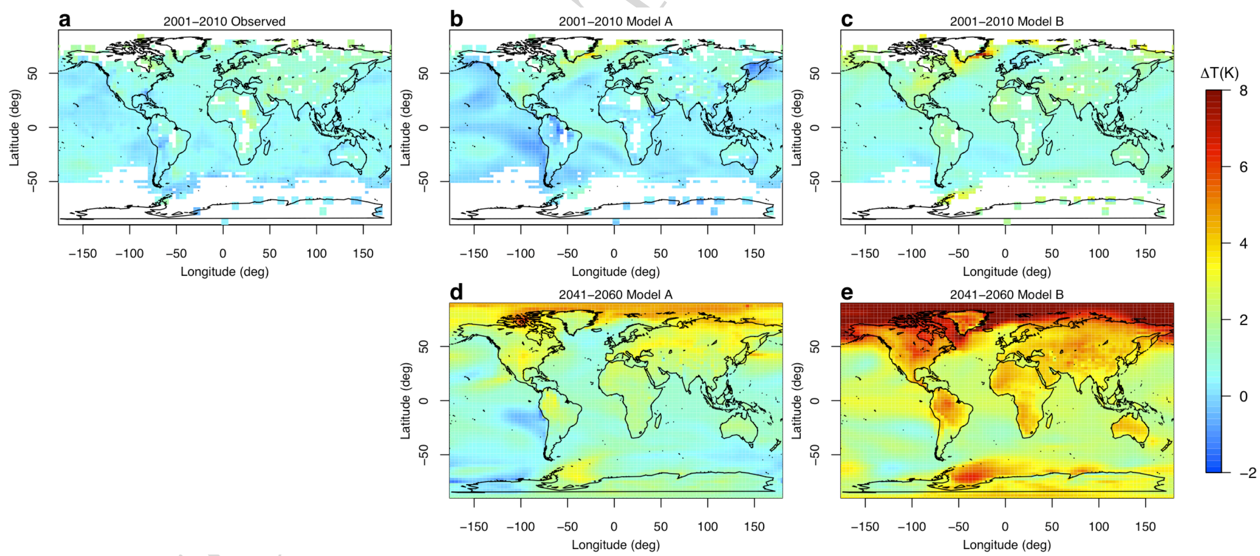

Figure 3: Surface temperature anomaly fields relative to 1961-1990 for 2001-2010 hindcast and 2041-2060 forecast for a low response ensemble member, A ($\Delta T_{2050} = 1.4K$) and high response ensemble member, B ($\Delta T_{2050} = 3K$). **a**, Observed 2001-2010 anomaly; **b, d** Model A anomaly for 2001-2010 and 2041-2060; **c, e** Model B anomaly. Both model versions are consistent with the surface temperature observations and are denoted by large labelled symbols in Fig. 2. White regions in **a** indicate missing data, defined as $> 60\%$ missing over 1961-1990 or 2001-2010. The same mask is applied in **b** and **c**.

Methods Summary

Model Simulations. HadCM3L consists of a of 3.75° longitude by 2.5° latitude atmosphere with interactive sulphur cycle coupled to a dynamical ocean of the same resolution[17]. Model physics parameters are perturbed through expert elicitation, and informed for atmospheric and sulphur cycle physics perturbations by results from the climate*prediction*.net slab model experiment[17] (Table SI 1,SI 2). Flux adjustments are calculated for 10 ocean configurations through a 200 year spin-up coupled to a standard atmosphere, and for each of 153 perturbed atmospheres[17], producing 1530 possible model versions. Model atmospheres have climate sensitivities ranging from 2-9K. Uncertainty in historical and future solar, volcanic forcing and anthropogenic sulphate emissions are accounted for in transient simulations (Fig. SI 1). After matching model simulations based on parameters and natural forcing scenarios, and averaging over initial condition ensembles there are 2752 matched transient-control pairs.

Goodness-of-fit calculation. We calculate a goodness-of-fit statistic (model error) for each simulation based on the spatio-temporal pattern of surface temperature from 1961-2010 as,

$$r_{\boldsymbol{\theta}}^2 = (\mathbf{y} - \mathbf{x}_{\boldsymbol{\theta}})^T \, \mathbf{C}_N^{-1} \, (\mathbf{y} - \mathbf{x}_{\boldsymbol{\theta}}) \, .$$

$\mathbf{y}$ represents observations, $\mathbf{x}_{\boldsymbol{\theta}}$ a transient-control pair of simulations corresponding to parameters $\boldsymbol{\theta}$, and $\mathbf{C}_N$ a covariance matrix describing variability in $\mathbf{y}$ and $\mathbf{x}_{\boldsymbol{\theta}}$ expected from internal variability, estimated from segments of CMIP-3 pre-industrial control runs[5] and a 1000 year HadCM3 control run respectively (Supplementary Information). We project all data onto the leading spatial EOFs of the HadCM3L ensemble of transient-control pairs, retaining over 90% of the ensemble variance. Uncertainty analysis is based on comparing a given $r_{\boldsymbol{\theta}}^2$ to the distribution expected from internal variability, using independent samples for estimating $\mathbf{C}_N$ and subsequent uncertainty analysis (Fig. SI 3). In Fig. 2 we display goodness-of-fit as a weighted mean squared error by normalising $r_{\boldsymbol{\theta}}^2$ by the number of degrees of freedom in $\mathbf{y}$ and $\mathbf{x}_{\boldsymbol{\theta}}$.

Acknowledgements We thank all participants in the climateprediction.net experiments, as well as the academic institutions and the individuals who have helped make the experiment possible, particularly David Anderson for developing the Berkeley Open Infrastructure for Network Computing. We also thank the Natural Environment Research Council, the EU FP6 WATCH and ENSEMBLES projects, the Oxford Martin School, the Smith School of Enterprise and the Environment and Microsoft Research for support and Jonathan Renouf and co-workers at the BBC for their documentaries explaining and promoting this experiment. D. J. R. was supported by a NERC PhD studentship with a CASE award from CEH

Wallingford.

Correspondence and requests for materials should be addressed to D. J. R. (e-mail: rowlands@atm.ox.ac.uk)

Competing interests statement the authors declare that they have no competing financial interests.

References

1. Stott, P. A. *et al.* Observational constraints on past attributable warming and predictions of future global warming. *J. Clim.* **19**, 3055–3069 (2006).

2. Harris, G. R. *et al.* Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Climate Dynamics* **27**, 357–375 (2006).

3. Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. Quantifying Uncertainties in Climate System Properties with the Use of Recent Climate Observations. *Science* **295**, 113–116 (2002).

4. Knutti, R., Stocker, T. F., Fortunat, J. & Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).

5. Meehl, G. A. *et al.* The WCRP CMIP3 multi-model dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.* **88**, 1383–1394 (2007).

6. Collins, M. *et al.* Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Climate Dynamics* **36**, 1737–1766 (2010).

7. Kiehl, J. Twentieth century climate model response and climate sensitivity. *Geophysical Res. Letter* **34** (2007). L22710.

8. Knutti, R. Why are climate models reproducing the observed global surface warming so well? *Geophysical Res. Letter* **35** (2008). L18704.

9. Huybers, P. Compensation between Model Feedbacks and Curtailment of Climate Sensitivity. *J. Clim.* **23**, 3009–3018 (2010).

10. Knutti, R. *et al.* A Review of Uncertainties in Global Temperature Projections over the Twenty-First Century. *J. Clim.* **21**, 2651–2663 (2008).

11. Forest, C. E., Stone, P. H. & Sokolov, A. P. Constraining climate model parameters from observed 20th century changes. *Tellus A* **60**, 911–920 (2008).

12. Friedlingstein, P. *et al.* Climate-carbon cycle feedback analysis: Results from the C^{4}MIP model intercomparison. *J. Clim.* **19**, 3337–3353 (2006).

13. Milly, P. C. D., Dunne, K. A. & Vecchia, V. Global pattern of trends in stream flow and water availability in a changing climate. *Nature* **428**, 347–350 (2005).

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

15. Murphy, J. M. *et al.* Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772 (2004).

16. Jackson, C. S., Sen, M. K., Huerta, G., Deng, Y. & Bowman, K. P. Error Reduction and Convergence in Climate Prediction. *J. Clim.* **21**, 6698–6709 (2008).

17. Frame, D. J. *et al.* The climate*prediction*.net BBC climate change experiment: design of the coupled model ensemble. *Phil. Trans. Roy. Soc. Lond. A* **367**, 855–870 (2009).

18. Nakicenovic, N. & Swart, R. *Special Report on Emissions Scenarios* (Cambridge University Press, 2000).

19. Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1950. *J. Geophys. Res.* **111** (2006).

20. Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. & Meehl, G. A. Challenges in combining projections from multiple climate models. *J. Clim.* **23**, 2739–2758 (2010).

21. Weigel, A. P., Knutti, R., Liniger, M. & Appenzeller, C. Risks of Model Weighting in Multimodel Climate Projections. *J. Clim.* **23**, 4175–4191 (2010).

22. Frame, D. J. *et al.* Constraining climate forecasts: The role of prior assumptions. *Geophysical Res. Letter* **32** (2005).

23. Easterling, D. R. & Wehner, M. F. Is the climate warming or cooling? *Geophysical Res. Letter* **36** (2009). L08706.

24. Solomon, S. *et al.* Contributions of Stratospheric Water Vapor to Decadal Changes in the Rate of Global Warming. *Science* **327**, 1219–1223 (2010).

25. Lockwood, M. Solar change and climate: an update in the light of the current exceptional solar minimum. *Phil. Trans. Roy. Soc. Lond. A* **466**, 303–329 (2010).

26. Stone, D. A. & Allen, M. R. Attribution of global surface warming without dynamical models. *Geophysical Res. Letter* **32** (2005). L18711.

27. Betts, R. A. *et al.* When could global warming reach 4^{o}C. *Phil. Trans. Roy. Soc. Lond. A* **369**, 67–84 (2011).

28. Desaii, S., Hulme, M., Lempert, R. & Pielke-Jr, R. Do We Need Better Predictions to Adapt to a Changing Climate? *EOS* **90**, 111–112 (2009).

29. Joshi, M. M., Webb, M. J., Maycock, A. C. & Collins, M. Stratospheric water vapour and high climate sensitivity in a version of the HadSM3 climate model. *Atmos. Chem. Phys. Discuss.* **10**, 6241–6255 (2010).

30. Holden, P. B. & Edwards, N. R. Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling. *Geophysical Res. Letter* **37** (2010). L21707.

Methods

Model Simulations. HadCM3L[31] is a version of the UK Met Office Unified Model using a horizontal grid of 3.75° longitude by 2.5° latitude with 19 levels in the vertical. The ocean resolution is the same as the atmosphere and consists of 20 vertical levels. The model contains an interactive sulphur cycle, simulating the direct and first indirect effects[32]. Ocean physics parameters are perturbed through expert elicitation[33], and atmospheric and sulphur cycle physics perturbations informed by results from the climate*prediction*.net slab model experiment[32,34], choosing between 2 and 4 values for each parameter (Table SI 1,SI 2). Atmospheric configurations are initially chosen to span a wide range of equilibrium climate sensitivities (2-9K) whilst still retaining an acceptable climatology, measured through the TOA flux imbalance relative to the standard physics settings ($\pm 10 W/m^2$)[34].

Typically AOGCMs require long spin-up periods in order to reach a stable equilibrium, and often when atmospheric and oceanic components are coupled together drifts can occur. A technique has been developed to allow a large number of drift-free coupled model simulations to be produced, with no need for a new ocean spin-up when the fast components of the model (atmosphere, land-surface scheme) are perturbed[17]. 10 versions of the HadCM3L ocean model coupled to the standard atmosphere are spun up for 200 years and necessary flux adjustments corresponding to the climate around 1920 calculated. Secondly, additional flux adjustments arising from atmospheric parameter perturbations are then calculated for each of 153 atmospheric versions, and added to the corresponding ocean flux adjustment, thus giving a total of 1530 different combinations of atmosphere and ocean physics. Each of the 1530 possible combinations ("model versions") with the associated total flux adjustment, are then run under a set of transient forcings from 1920-2080 and also under control forcing for the same length of time in initial condition ensembles.

Uncertainty in historical natural forcing is represented through 5 solar and 5 volcanic scenarios, and in the future through 3 solar and 10 volcanic scenarios (Fig. SI 1b,d). We use a set of scalings on historical and future (SRES A1B) sulphate emissions generating model sulphur cycle responses consistent with current estimates of uncertainty[35] (Fig. SI 1c). SRES A1B[18] represents a mid-range emissions scenario and given the limited impact of emissions scenario by 2050[36] is expected to produce qualitatively similar results to the newer RCP 4.5 mid-range scenario[37].

Simulations are run on computers volunteered by the general public: in total 9745 simulations returned complete data. Given bandwidth and storage constraints in the distributed computing environment, each simulation returns "trickle" files on a yearly basis, consisting of monthly time-series averaged over 61 regions over the globe, and upload files every 10 years containing seasonally averaged full field output.

We restrict our analysis to the surface temperature data focussing on 22 Giorgi land regions[38] and 6 major ocean basis for our comparison with observations (Table SI 3). Matched "transient minus control" pairs are used to remove any unforced drifts due to residual energy imbalances in the coupling process[17].

Data Preparation. Of the 9745 complete simulations there are 1656 controls and 8089 transients. Basic quality control on the model simulations is applied. Model versions with absolute global mean drifts in the control climate larger than 0.4K/century are flagged, indicating the flux adjustment has not eliminated unforced drifts. Transient simulations are matched based on their parameters and natural forcing scenario. Initial condition ensemble averages are taken where possible to reduce noise in the model simulations. Controls are prepared identically, and matched to corresponding transients through the model parameters, giving a total of 2752 distinct transient-control pairs. A control simulation can be matched to many transients given the separation by natural forcing or anthropogenic sulphate scaling.

The 2752 transient-control pairs contain 809 of the original 1530 possible model versions. Each transient-control pair is expressed as an anomaly from the 1961-1990 mean in each region. Observations, HadCRUT3[19] for land and HadSST2[39] for ocean, CMIP-3[5] and QUMP[6] simulations under the A1B scenario and CMIP-3 pre-industrial control simulations are prepared identically (Table SI 4). Finally, all data is temporally averaged to 5 year mean resolution to reduce the impact of internal variability. For simplicity, coverage is assumed complete within Giorgi regions in this analysis of the model output: this introduces only small errors since the regions used have a high observational coverage ($> 90\%$) over the 1961-2010 period considered (Fig. 3a).

Goodness-of-fit calculation. We calculate a goodness-of-fit statistic (model error) based on the spatio-temporal pattern of surface temperature from 1961-2010 as,

$$r_{\boldsymbol{\theta}}^2 = \left(\mathbf{y} - \mathbf{x}_{\boldsymbol{\theta}}\right)^T \mathbf{C}_N^{-1} \left(\mathbf{y} - \mathbf{x}_{\boldsymbol{\theta}}\right),$$

where $\mathbf{y}$ represents observations, $\mathbf{x}_{\boldsymbol{\theta}}$ a transient-control pair of simulations corresponding to parameters $\boldsymbol{\theta}$, and $\mathbf{C}_N$ a covariance matrix which weights errors corresponding to the expected variability in components of $\mathbf{y}$ and $\mathbf{x}_{\boldsymbol{\theta}}$ arising from internal climate variability. Observations cannot easily be used to estimate $\mathbf{C}_N$ without simplifying assumptions, and so segments of pre-industrial control simulations are used as is standard practice[40]. We use pre-industrial control simulations from all available CMIP-3 models to account for variability in $\mathbf{y}$ thus allowing for model uncertainty in the covariance estimation[41], and a 1000 year HadCM3 control run[42] to characterise variability in $\mathbf{x}_{\boldsymbol{\theta}}$. We find little sensitivity in the results to scaling the variability associated with $\mathbf{y}$ over a wide range (Fig. SI 10).

Estimates of variability from AOGCMs are most reliable on large spatial scales, so we focus on the leading

Empirical Orthogonal Functions (EOFs) of the BBC CCE over 1961-1990, the first 3 of which explain over 90% of the spatial variance across the ensemble. The exact choice of truncation does not significantly impact results when using a regularized covariance estimate[43], and using a separate physically-based dimension reduction technique does not change our conclusions (Fig. SI 8).

For a given confidence level, we compare $r^2_{\boldsymbol{\theta}}$ to the corresponding percentile of the distribution of r^2 arising from estimates of internal variability alone using the pre-industrial control segments. A schematic of the analysis is shown in Fig. SI 3. We use an independent set of control segments to $\mathbf{C}_N$ to remove the small sample size bias[40]. This tests the null hypothesis that the model and observations come from the same distribution and rejects the model simulation if $r^2_{\boldsymbol{\theta}}$ is too large. In Fig. 2 we display goodness-of-fit as a weighted mean squared error by normalising $r^2_{\boldsymbol{\theta}}$ by the number of degrees of freedom in $\mathbf{y}$ and $\mathbf{x}_{\boldsymbol{\theta}}$.

Methods References

31. Jones, C. D. & Palmer, J. R. Spinup methods for HadCM3L. Tech. Rep. CRTN 84, Hadley Centre for Climate Prediction and Research (1998).

32. Ackerley, D., Highwood, E. J. & Frame, D. J. Quantifying the effects of perturbing physics of an interactive sulfur scheme using an ensemble of GCMs on the climateprediction.net platform. *J. Geophys. Res.* **114** (2009). D01203.

33. Collins, M. *et al.* The Sensitivity of the Rate of Transient Climate Change to Ocean Physics Perturbations. *J. Clim.* **20**, 2315–2320 (2007).

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

35. Textor, C. *et al.* Analysis and quantification of the diversities of aerosol life cycles within AeroCom. *Atmos. Chem. Phys.* **6**, 1777–1813 (2006).

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

37. Moss, R. *et al.* Towards New Scenarios For Analysis Of Emissions, Climate Change, Impacts and Response Strategies. Tech. Rep., IPCC (2008).

38. Giorgi, F. & Francisco, R. Evaluating uncertainties in the prediction of regional climate change. *Geophysical Res. Letter* **27**, 1295–1298 (2000).

39. Rayner, N. A. *et al.* Improved Analyses of Changes and Uncertainties in the Sea Surface Temperature Measured In Situ since the Mid-Nineteenth Century: The HadSST2 Dataset. *J. Clim.* **19**, 446–469 (2006).

40. Allen, M. R. & Tett, S. F. B. Checking for model consistency in optimal fingerprinting. *Climate Dynamics* **15**, 419–434 (1999).

41. Gillett, N. P. *et al.* Detecting anthropogenic influence with a multi-model ensemble. *Geophysical Res. Letter* **29** (2002).

42. Collins, M., Tett, S. F. B. & Cooper, C. The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics* **17**, 61–81 (2001).

43. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88**, 365–411 (2004).