Climate projections: Past performance no guarantee of future skill?

C. Reifen[1] and R. Toumi[1]

Received 16 April 2009; revised 28 May 2009; accepted 8 June 2009; published 7 July 2009.

[1] The principle of selecting climate models based on their agreement with observations has been tested for surface temperature using 17 of the IPCC AR4 models. Those models simulating global mean, Siberian and European 20th Century surface temperature with a lower error than the total ensemble for one period on average do not do so for a subsequent period. Error in the ensemble mean decreases systematically with ensemble size, N, and for a random selection as approximately $1/N^{\alpha}$, where α lies between 0.6 and 1. This is larger than the exponent of a random sample (α = 0.5) and appears to be an indicator of systematic bias in the model simulations. There is no evidence that any subset of models delivers significant improvement in prediction accuracy compared to the total ensemble. **Citation:** Reifen, C., and R. Toumi (2009), Climate projections: Past performance no guarantee of future skill?, *Geophys. Res. Lett.*, *36*, L13704, doi:10.1029/2009GL038082.

1. Introduction

[2] With the ever increasing number of models, the question arises of how to make a best estimate prediction of future temperature change. The Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) combines the results of the available models to form an ensemble average, giving all models equal weight. Other studies argue in favor of treating some models as more reliable than others [*Shukla et al.*, 2006; *Giorgi and Mearns*, 2002]. However, determining which models, if any, are superior is not straightforward. The IPCC comments: "What does the accuracy of a climate model's simulation of past or contemporary climate say about the accuracy of its projections of climate change? This question is just beginning to be addressed..." [*Intergovernmental Panel on Climate Change*, 2007, p. 594]. One key assumption, on which the principle of performance-based selection rests, is that a model which performs better in one time period will continue to perform better in the future. This has been studied in terms of pattern-scaling using the "perfect model assumption" [*Whetton et al.*, 2007]. We examine the question in an observational context for temperature here for the first time. We will also quantify the effect of ensemble size on the global mean, Siberian and European temperature error.

[3] The principle of averaging results from different models to form a multi-model ensemble prediction also has potential problems, since models share biases and there is no guarantee that their errors will neatly cancel out. For this reason groups of models thus combined have been termed "ensembles of opportunity" [*Piani et al.*, 2005]. Various studies have showed that multi-model ensembles produce more accurate results than single models [*Kiktev et al.*, 2007; *Mullen and Buizza*, 2002]. Our examination of ensemble performance aims to address the question in the context of the current generation of climate models.

2. Data

[4] We examine model simulations of 20th Century climate. The data is part of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) [*IPCC*, 2007] and is available from the Coupled Model Intercomparison Project (CMIP, see https://esg.llnl.gov:8443). The 17 models included in this analysis are those that have run the A2 emissions scenario, so they form the ensemble for one scenario (listed alphabetically: BCCR-BCM2.0, CCCMA-CGCM3.1(T47), CNRM-CM3, CSIRO-MK3.0, GFDL-CM2.0, GFDL-CM2.1, GISS-ER, INM-CM3.0, IPSL-CM4, MIROC3.2(medres), MIUB-ECHO-G, MPI-ECHAM5, MRI-CGCM2.3.2, NCAR-CCSM3, NCAR-PCM, UKMO-HadCM3, UKMO-HadGEM1; see http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php). Models are evaluated against the HadCRUT3 5° × 5° gridded surface air temperature observations [*Brohan et al.*, 2006]. All modeled and observed temperatures used here are anomalies with respect to the 1961–1990 average.

3. Methodology

[5] The key question is: if a sub-group of the 17 models outperforms the total ensemble average in one time period, will that sub-group continue to outperform? We ranked the models in order of lowest error in the 10, 20 and 30-year mean compared with observations over the period 1900–1999, using moving averages. Where errors are described as "gridded", the spatially varying root mean square error (RMSE) is calculated across the grid points in the region. In the global mean case, the error is simply the 10, 20 or 30-year mean model bias. The model output is bilinearly interpolated onto the 5° × 5° observational grid. The N models with best performance in the "selection" period (e.g. 1900–1919) are combined to produce a multi-model ensemble mean time series for the next non-overlapping "test" period (e.g. 1920–1939). The number of models included, N, is varied from 1 to 16 to examine the whole range of ensemble sizes. The total ensemble mean is also calculated for each period, to see whether or not a selected model ensemble can outperform the average of all 17 models. This process is then repeated with the next time block (e.g. 1901–1920 selection period, 1921–1940 test period). It is important to note that for each time block there is a turnover of models. We are not keeping the same models

[1] Department of Physics, Imperial College London, London, UK.

0094-8276/09/2009GL038082

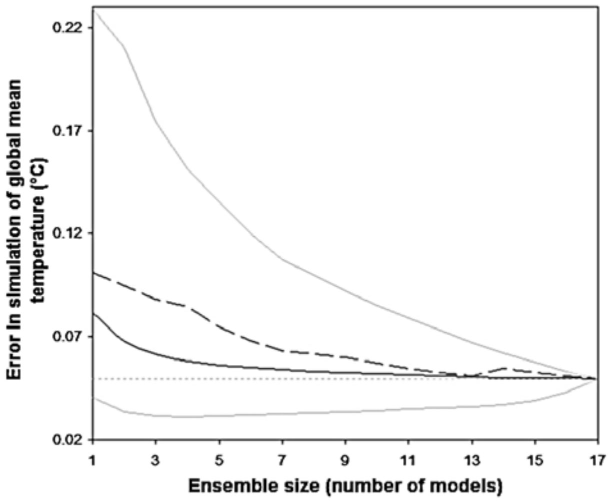

Figure 1. RMS error of predicted global 20-year mean surface temperature for the 17-model ensemble (grey dotted), best selected ensembles of all sizes (black dashed) and the mean of all possible combinations of models (or random selection) for each ensemble size (solid black); solid grey lines show the combinations of ensembles with minimum error (bottom line) and maximum error (top line); the grey dotted line does not correspond to the different values on the "Ensemble size" axis, but is intended for ease of reference against the other curves.

in an ensemble, so that the average performance for a given ensemble size is the best possible based on the range of models available.

4. Results

[6] Figure 1 shows how the error in global mean surface temperature of progressively larger selected ensembles (black dashed line) approaches the error of the whole 17-model ensemble (grey dotted line). The solid black curve is intended to represent the average random selection of how the error decreases with ensemble size for this particular spatial scale and time period. No performance-based selection is employed in computing the random selection since it is intended to represent the average behavior expected from ensembles of models selected at random. Thus the error of an ensemble of size of 1 is the mean of each individual model's error (an average of 17 values) and at an ensemble of 17 it is the grey dotted line representing the whole ensemble. For an ensemble of 8, the random selection is the average of the errors of 24310 different ensembles, all possible ways of selecting 8 models out of the total 17.

[7] These results advocate the use of multi-model ensembles as preferable to choosing a single model or a smaller ensemble of selected models when making best estimate projections. Whether models are selected according to their past performance or at random, the prediction error decreases systematically with the inclusion of more models in the ensemble. The other striking feature is that, while the error decreases steeply at smaller ensemble sizes, the improvement in accuracy achieved by adding more models wanes quickly thereafter. We speculate that the initial rapid decrease in error is probably an artifact due to poor sampling and high variance of a single model run. In this global case, performance-based model selection offers no improvement, in fact the errors of model groups chosen for their previous good performance are actually larger than the corresponding average random selection. It is also worth noting that the random selection curve is the mean of a wide range of ensemble behaviors. The solid grey lines in Figure 1 represent the ensembles with minimum error (bottom curve, the best possible ensemble) and maximum error (top curve, the worst ensemble). This illustrates the risks associated with smaller ensemble size. Interestingly, the average random selection error is closer to the minimum error than to the maximum, suggesting a negative skew in the error distribution. It is also notable how close the total model ensemble error (grey dotted line) lies to the minimum error curve. Simply averaging all models achieves nearly as high an accuracy as even the best performing subsets of the available model group.

[8] The analysis was repeated with 10 and 30-year averages, as well as using other metrics (correlation and linear trend) with the same result: selecting models for "best past performance" does not appear to convey any future benefits. For example, using 30-year averages the lowest RMSE of any selected ensemble is 0.03°C (with an ensemble of 12 models), which is the same as the error of the total 17-model ensemble. Using the 10-year smoothed linear decadal trend as the metric, the total ensemble has an average error of 0.09°C/decade. Selected ensembles with sizes of 13 models and above have the same error as the total ensemble, but all smaller selected ensembles have larger errors in the decadal trend. As in the 20-year case, the ensembles are selected based on performance in one 10 or 30-year time block and the errors of the selected ensembles are evaluated in the consecutive (non-overlapping) time block.

[9] To test the global analysis regionally we chose Siberia (50–70°North, 60–130°East) and Europe (35–60°North, 0–45°East) as areas of interest. The selection process was carried out as above but using the area-averaged surface temperature and gridded error, which is the average error of the grid points in the region. For a particular gridpoint and time period to be included, we require 10 years of observational data, not necessarily continuous, in each 20-year mean period. We have also obtained similar results with different data requirements. Figure 2 shows the error in gridded Siberian and European 20-year mean surface temperature of selected ensembles (black dashed line) compared with that of the whole 17-model ensemble (grey dotted line). The solid black curve again represents the average random selection. The results are similar to those for global mean temperature in Figure 1. As would be expected, the regional errors are larger than the global mean.

[10] In the Siberian case, selected ensembles of 2–5 models do slightly outperform the total 17-model ensemble on average, but in fact they only have lower error in 62% of the test periods. In the European case the error of the selected ensembles lies above both the total 17-model ensemble error and the average random selection curve. The region-averaged Siberian and European results also show selected ensembles rarely outperforming the total ensemble and never by a substantial amount. For example, the error of the whole ensemble in simulating Siberian area-

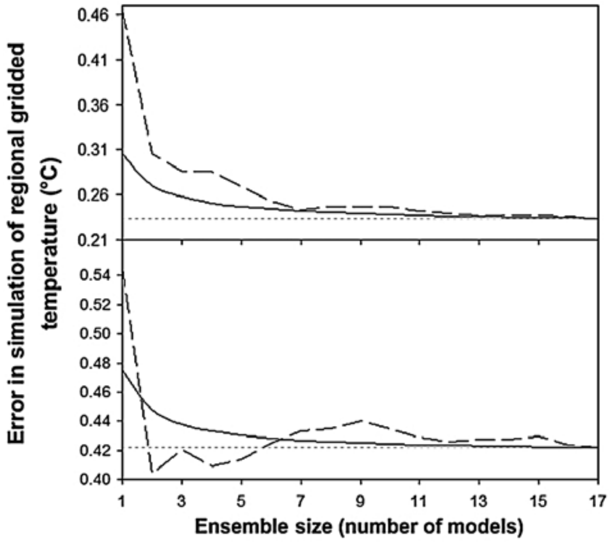

Figure 2. RMS error of gridded (top) European and (bottom) Siberian 20-year mean surface temperature for the 17-model ensemble (grey dotted), best selected ensembles of all sizes (black dashed) and the mean of all possible combinations of models (or random selection) for each ensemble size (solid black); the combinations of ensembles with minimum error (bottom) and maximum error (top) are in solid grey; the grey dotted line does not correspond to the different values on the "Ensemble size" axis, but is intended for ease of reference against the other curves.

averaged 20-year mean temperature is 0.18°C, which is the same as the lowest error of any of the selected ensembles. Analysis of other metrics (correlation and linear trend) led to the same conclusion.

[11] Not all models include solar and volcanic forcings, which are particularly important over the first half of the 20th Century. Five out of the seventeen models used here omit one or both of these forcings (BCCR-BCM2.0, CNRM-CM3, CSIRO-MK3.0, IPSL-CM4, MPI-ECHAM5). We repeated the analysis excluding these models, with similar results. In the 20-year average global mean and European gridded cases, the selected ensemble error converges earlier on the error of the whole ensemble, but never falls below the total ensemble error. In the 20-year average Siberian gridded case, selected ensembles of 2 models have an average RMSE of 0.41°C, which is lower than the total ensemble RMSE of 0.43°C, but the selected ensembles only outperform in 60% of the test periods. For all other ensemble sizes, the selected ensembles have higher RMSE than the total ensemble.

[12] Since the greenhouse gas signal becomes increasingly dominant, it could be argued that recent model skill is more important. Using the eleven 1980–1999 test periods in the global 10-year mean case, the error of the whole ensemble is 0.06°C and the lowest error of any selected ensemble is 0.05°C. The selected ensemble only outperforms the whole ensemble in 60% of the test periods. In the Siberian gridded analysis the mean RMSE of the whole ensemble is 0.48°C (0.51°C) over the 1980–1999 (1910–1999) period, which in both cases is the same as the lowest error of any of the selected ensembles.

[13] Rather than choosing the "best *N*", it may be more appropriate to select models that agree with observations within acceptable margins. This method has been tested and does not significantly change the results. For example, we selected only those models that lie within the observational HadCRUT3 uncertainty range plus one standard deviation of all the models. In the 20-year global mean case, the error of the whole ensemble is 0.05°C and that of the selected ensemble is 0.06°C. The selected ensemble varies between 6 and 16 models (average 9). The error is equivalent to that of selecting the 9 best models throughout. Error margin and ranking selections are very similar and do not seem to offer any advantage over the whole ensemble.

[14] Even in cases where the selected ensemble delivers very low errors in the selection time period, these improvements are not always propagated forward to the test period. There is a lack of persistence in the relative skill of the models, which can be illustrated by looking at the turnover in membership of the selected ensemble. Turnover is defined as follows: with the models ranked on performance, the percentage of models in the top *N* in the first time block which drop out of the top *N* in the next non-overlapping time block. Figure 3 shows the mean turnover (averaged over all time blocks) against all possible choices of *N*, from choosing just the one best model to excluding just the one worst model (*N* = 16). Turnover for Europe and Siberia (corresponding to the results in Figure 2) is shown as well as for the global mean (corresponding to Figure 1). As expected, if a choice is made to select just the one best model, there is minimal chance that it will continue to be the best model. Even larger ensemble sizes exhibit high turnover, in spite of being limited by the relatively small number of 17 possible models. Different averaging periods produce similar results, for example the mean turnover in an ensemble of 8 models for the global mean case is 62% using 20-year averages, compared to 59% (46%) using 30-year (10-year) averages.

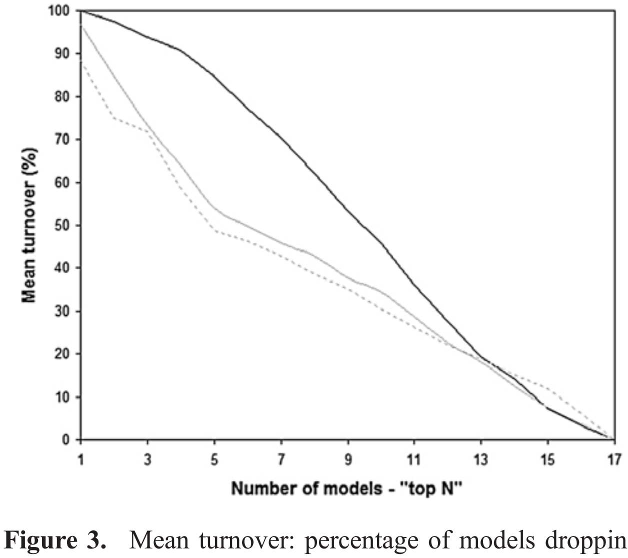

Figure 3. Mean turnover: percentage of models dropping out of the top *N* from one time block to another, plotted against all values of *N* from 1 to 17; models are ranked according to their error in simulating 20-year average global mean (solid black), Siberian gridded (grey dotted) and European gridded (solid grey) surface temperature over the 20th century.

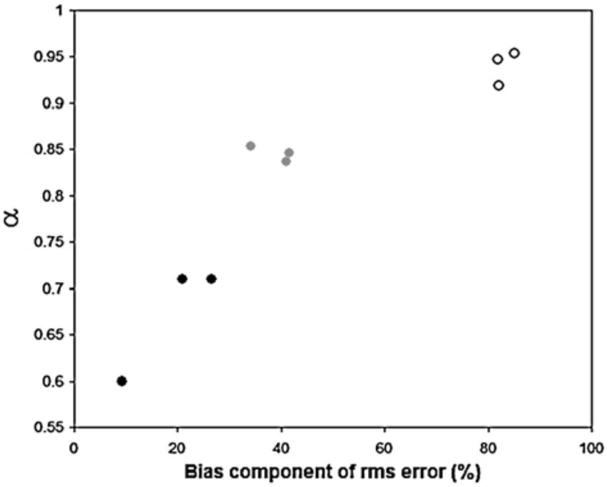

Figure 4. Variation of the power law exponent, α, with the mean bias component of the RMS error of the 17-model ensemble (expressed as a percentage of the total RMS error) for global mean results (solid black circles), European gridded results (solid grey circles) and Siberian gridded results (open black circles); errors are in the anomaly relative to the 1961–1990 average; the 3 values for each region are the 10, 20 and 30-year mean results.

[15] The random selection error curves for the globe, Siberia and Europe approach the 17-model ensemble as a power law: $1/N^{\alpha}$, where α lies between 0.6 and 1. This power law exponent can be used as a quantification of the increased accuracy as more models are added to the ensemble. The largest improvements in accuracy occur at the lower end of ensemble size, with reductions in error becoming smaller as more models are added. Stochastically generated time series, simulating Gaussian white noise, were analyzed as above. This showed that the exponent of the power law decrease in RMSE tends to $\alpha = 0.5$ for large samples, which is lower than the 0.6–1 range seen in the results of the model analysis. We suggest that exponents greater than 0.5 are evidence of intercorrelation between the errors of different models and therefore of model interdependence. Figure 4 shows that in general α increases with the percentage of RMSE resulting from systematic bias in the total 17-model ensemble time series. This tendency is also seen in stochastically generated sets with introduced bias. The presence of systematic bias means that there will be correlations between the errors of different models with respect to observations. This result therefore corroborates the argument that where errors in model simulations are less correlated they produce a lower exponent. The global results are quite close to the random value of $\alpha = 0.5$, particularly the 10-year mean which has $\alpha = 0.6$ and a bias of just 9%. However, the European and Siberian exponents are much larger, approaching $\alpha = 1$ in the Siberian case, indicating that concerns about lack of model independence [*Tebaldi and Knutti*, 2007] may be well-founded on regional scales.

5. Discussion

[16] In our analysis there is no evidence of future prediction skill delivered by past performance-based model selection. There seems to be little persistence in relative model skill, as illustrated by the percentage turnover in Figure 3. We speculate that the cause of this behavior is the non-stationarity of climate feedback strengths. Models that respond accurately in one period are likely to have the correct feedback strength at that time. However, the feedback strength and forcing is not stationary, favoring no particular model or groups of models consistently. For example, one could imagine that in certain time periods the sea-ice albedo feedback is more important favoring those models that simulate sea-ice well. In another period, El Niño may be the dominant mode, favoring those models that capture tropical climate better. On average all models have a significant signal to contribute.

[17] Ideally we would want to test models which have been developed independently of observations. We recognize the importance of the post 1970 climate shift period and it would have been instructive to test this period with models having no prior knowledge of it. However, this problem also applies to any climate projections using a subset of models based on 20th Century performance and it is this approach we are testing here. We are not implying that comparisons against observations are not important in model validation. Good agreement with past climate builds confidence in the reliability of a model's future projections. Our analysis only examines selection based on models' ability to replicate a mean anomaly over a historic time period. There are other criteria that could be used and would be worth investigating.

[18] Using the current generation of models, the best estimate of climate change is unlikely to benefit substantially by increasing the number of AOGCMs. However, we do not know which feedbacks will dominate in the future and the inclusion of the largest possible number of models could increase the range of predictions, which may be more useful than the best estimate, and will also reduce the standard error of the mean projection. We therefore conclude that the multi-model ensemble mean of all available AR4 models provides the most accurate basis for making best estimate projections of future climate change. The common investment advice that "past performance is no guarantee of future returns" and to "own a portfolio" appears also to be relevant to climate projections.

[19] **Acknowledgments.** We thank NERC and BP for funding this work, CRU for the gridded observational data, the international modeling groups for providing their data for analysis, the PCMDI for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modeling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC multi-model data archive at LLNL is supported by the Office of Science, U.S. Department of Energy.

References

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548.

Giorgi, F., and L. Mearns (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.

Intergovernmental Panel on Climate Change (IPCC) (2007), *Climate Change 2007: The Scientific Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, Cambridge, U. K.

Kiktev, D., J. Caesar, L. V. Alexander, H. Shiogama, and M. Collier (2007), Comparison of observed and multimodeled trends in annual extremes of temperature and precipitation, *Geophys. Res. Lett.*, *34*, L10702, doi:10.1029/2007GL029539.

Mullen, S. L., and R. Buizza (2002), The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System, *Weather Forecast.*, *17*, 173–191.

Piani, C., D. J. Frame, D. A. Stainforth, and M. R. Allen (2005), Constraints on climate change from a multi-thousand member ensemble of simulations, *Geophys. Res. Lett.*, *32*, L23825, doi:10.1029/2005GL024452.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino (2006), Climate model fidelity and projections of climate change, *Geophys. Res. Lett.*, *33*, L07702, doi:10.1029/2005GL025579.

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

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady (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:10.1029/2007GL030025.

C. Reifen and R. Toumi, Department of Physics, Imperial College London, London SW7 2BW, UK. (catherine.reifen02@imperial.ac.uk)