Interannual Temperature Predictions using the CMIP3 multi-model ensemble mean

Thomas Laepple[1], Stephen Jewson[2], Katie Coughlin[2]

[1]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

[2]Risk Management Solutions, London

thomas.laepple@awi.de

We present a simple method to make multi-year surface temperature forecasts using the climate change simulations of the CMIP3 database prepared for the IPCC AR4 report. By calibrating the multi-model ensemble mean with current observations, we are able to make skillful interannual forecasts of mean temperatures. The method is validated using extensive hindcast experiments and is shown to perform favorably compared to a recent forecast method based on a global circulation model with assimilated initial conditions. Five year forecasts for the global mean temperature, the Northern Hemispheric mean temperature and the summer sea surface temperatures (SSTs) in the main development region for hurricanes (MDR) are presented.

1. Introduction

The latest report of the Intergovernmental Panel on Climate Change (IPCC) (Solomon 2007) presented long-term projections of climate change into the next century. It was emphasized that most of the observed warming over the past 50 years is attributable to human activities and that the climate will likely continue to warm. Whereas the projections of the report are made on the century scale, industry and policy makers are often interested in a mid-term perspective of 1-10 years to plan their actions. Therefore there is also great interest in multi-year forecasts for the climate system.

Global seasonal-timescale climate predictions based on coupled ocean-atmosphere models are now operational in a large number of meteorological institutes but interannual forecasts using these models are still in development (e.g. Palmer, Alessandri et al. 2004 and references therein). Recently Smith et al. (2007) presented, for the first time, a mid-term, interannual global forecast which accounts for the effect of external forcing as well as internal variability. This decadal climate prediction system (DEPRESYS) is based on a coupled global climate model and takes into account the observed state of the atmosphere and ocean in order to predict the internal variability out to decadal time-scales. However, because this kind of forecast system is still developmental, the skill of the forecast needs to be weighed against the large technical and computing effort needed to implement such a system.

We present a very simple approach for interannual temperature forecasts using the existing output from the large ensemble of coupled ocean-atmosphere models which participated in the Coupled Model Intercomparison Project (CMIP3). By calibrating the model output with observed data, we use both the skill of the complex models in forecasting the anthropogenic contribution to changing temperatures and the skill of persistence, which is inherent in the temperature timeseries. Using this precompiled source of information, with appropriate bias corrections, we are able to make skillful interannual temperature predictions and we suggest that this simple prediction technique serve as a benchmark for future prediction experiments.

We demonstrate our prediction technique on three temperature indices: the annual global mean surface temperature (SAT) which exhibits very small interannual variability due to the large area mean, the Northern Hemispheric mean SAT, and the summer sea surface temperature (SST) in the main development region (MDR). SSTs in this Atlantic region exhibit very strong interannual to multi-decadal variability and are of special interest due to the possible connection to hurricane frequency and intensity (e.g. Goldenberg, Landsea et al. 2001; Emanuel 2005). Forecasts of these indices are given for a five-year outlook and the skill of the interannual forecasts is compared to Smith et al. [2007].

2. Data

We use the annual mean Land-Ocean Temperature anomaly Index for the Northern Hemispheric (NH) and Global mean Temperature (GL) provided by NASA GISS

(Hansen, Ruedy et al. 2006) (available at http://data.giss.nasa.gov/gistemp/). The HADISST dataset (Rayner, Parker et al. 2003) is used to extract the MDR SST index (15-70W,10-20N, JAS mean). We use an anomaly relative to 1951-1980. The three timeseries are shown in Figure 1a-c.

The model data consists of gridded global monthly SAT and SST from the World Climate Research Programme's Coupled Model Intercomparison Project multi-model dataset (CMIP3) (available at http://www-pcmdi.llnl.gov).

We extract mean temperatures over the seasons and regions which correspond to the observational data described above to create analogous time series for each model run. The historical scenario 20C3M as well as the future IPCC-scenarios SRESA1B, SRESA2 and SRESB1 are used.

3. Forecast method

We divide the models into a set which includes historical volcanic forcing and a set without volcanic forcing. As the volcanic forcing has a strong impact on the temperature timeseries, especially on the MDR SST (Santer, Wigley et al. 2006), but is not predictable in the future, we only use the non-volcanic models in this study to allow for fair hindcasts. The historical 20C3M simulations are merged with the future simulations. The concatenated simulations are then treated as continuous timeseries for the rest of the study. The models BCC-CM1 as well as the SRESB1, CSIRO-Mk3.0 runs were removed from the set to avoid discontinuities in 2000 as they did not restart from the last year of the 20C3M run.

For the next decade, the differences in the forcing of the scenarios are small (Zwiers 2002) so, in order to increase the size of our ensemble, we have included runs from all three. By taking the mean over all the ensemble members of the models and over these three scenarios we are able to remove most of the internal variability of the models. The resulting non-volcanic ensemble mean timeseries are shown in Figure 1a-c together with the observed timeseries.

In order to create a prediction of a temperature timeseries for the years n+1 onwards, a bias correction is needed to shift the ensemble mean to the current state of the observed temperatures. The current state is estimated using a number of years, N, before the current date, n. The correction then involves subtracting an average of the ensemble mean values over these years (n-N,n-N+1, …, n) and adding an average of the observed values over these years.

Applying this bias correction, we predict future temperature values from simulated values for the years n+1 onwards. We call this IPCC/CMIP3 ensemble based method IENS. The IENS approach is similar to the reference method NOASSIM from Smith et al. [2007], but the use of our optimized N year baseline takes into account slow natural variability.

As reference predictions, we provide an optimal persistence forecast which is the mean of the N years before the current date (we call this FLAT), and a simple persistence estimate which is the value of the year before the forecast (we call this PERSISTENCE). By construction of the FLAT forecast, the IENS forecast will have higher skill if, on average, the trend of the ensemble mean is realistic. We note here that a linear trend prediction, modeled using an optimized window length for the trend fit, was initially included for comparison. Although not shown here, the skill of this forecast was always less than that of the IENS method, and often less than for the FLAT method.

Obviously the forecasts IENS and FLAT depend on the calibration window length, N. The optimal N depends on the properties of the timeseries as well as on the lead time and is determined by hindcasting on the historical data where N is defined to be the number of years which minimizes the root mean squared error (RMSE). Figure 2 shows the dependence of the RMSE of 5-year mean hindcasts on N based on hindcasts from 1930-2006.

In terms of forecast error, there is an optimal calibration window, which in this case is seven years, for all of the IENS hindcasts. The RMSE of the IENS methods are lower than the RMSE of the FLAT method which shows that the CMIP3 ensemble mean adds skill to the forecast. How can we explain the shape of the calibration window length dependence? For very short calibration windows, the mean state is not well estimated and its large variance dominates the RMSE. Therefore, as the calibration window increases the RMSE decreases approximately as the standard error of the mean decreases (1/sqrt[N]). For long calibration periods, biases between the observations and the model mean, due to natural variability or structural errors of the models, become important and contribute to an increasing RMSE. A balance between these effects gives the minima seen in Figure 2.

4. Validation method

To compare prediction methods, we use the RMSE of hindcast experiments. For each hindcast, the window length for the FLAT and the IENS method are re-estimated using all data except an interval of 10 years surrounding the years to be hindcast. This is done to minimize the artificial inflation of forecast skill which occurs when the window length, N, is estimated using the same data as is used to validate the forecast.
Because there is a limited hindcast period, we also supply the 90% bootstrapping confidence intervals to estimate the uncertainty of the RMSE. These confidence intervals are derived by randomly sampling (with replacement) m hindcast errors where m is the total number of hindcasts (e.g. for the quinquennial forecast, m = 73). This is repeated 10,000 times and a RMSE is estimated each time to derive a distribution.

This will not, however, account for systematic errors that may be found in our estimate of the prediction error. There are some reasons why the hindcast RMSE may be a conservative estimate of the forecast RMSE:

1.) If there are no volcanoes during the forecast period, the error may be smaller than estimated since the hindcast is performed over past periods which did include volcanoes.

2.) The mean of three scenarios is used for the forecast, but there is only one scenario for most hindcast years. Therefore, the residual of the internal variability is smaller for years after 2000, which might reduce the forecast error.

3.) We perform the validation on all available years (1930-2006) to represent the natural variability. However, one can argue that the higher ratio, of externally forced change to natural variability, in recent years will reduce the future error of the IENS approach.

There are also reasons why our hindcast RMSE may be optimistic:

1.) The uncertainty of the future model forcing scenario is only represented by the last years of the hindcast experiment.

2.) Some of the model results may be tuned to the observational period causing the IENS hindcasts to be closer to the observations and artificially lowering the RMSE.

5. Results of validation and forecasts

The estimated RMSE for the different methods are compared in Figure 1d. These RMSEs are slightly higher than the minimum RMSE in Figure 1a-c since these errors also include the uncertainty in the window length estimation. Figure 1d shows that the IENS forecast is generally more accurate than the reference methods, FLAT and PERSISTENCE. However, the 90% bootstrap confidence interval shown by the error bars on the IENS value indicates that the ensemble mean forecast is significantly better than the PERSISTENCE forecast but not necessarily better than the FLAT forecast for the MDR SSTs. This is understandable if much of our prediction skill comes from the bias-correction, or estimate of the current state. The added skill due to the anthropogenic changes modeled by the ensemble mean is most obvious in the global mean and NH mean temperature where natural variability is small due to the larger spatial averaging. This result is consistent with the results from (Lee, Zwiers et al. 2006), who found decadal climate prediction skill of the global mean temperature due to changes in anthropogenic forcing.

Next we compare the skill of our method and the method of Smith et al. (2007). They use the HadCM3 model, with assimilated inital conditions, to predict temperatures out to 9 years.

Figure 3 shows the RMSE of annual mean global temperature forecasts using the IENS, FLAT and PERSISTENCE method for lead times from 1-9 years. The hindcasts are based on the period 1939-2006. 1939 is chosen as the initial year for the hindcasts as we test window lengths up to 30 years. The IENS method shows the most skill for all lead times and all three forecast methods show a decrease in skill for longer lead times. The difference between FLAT and PERSISTENCE RMSE decreases with lead time whereas the difference between IENS and FLAT increases with lead time. The reason for this is that when the bias dominates, for the FLAT and PERSISTENCE models, the better estimate of the mean state becomes less important.

Since IENS predicts a realistic trend on average, the increase in RMSE with lead time is slower. In Figure 3b we show the same results using the hindcast years 1983-2004, as in Smith et al. [2007]. It can therefore be directly compared to Figure 1a) of Smith et al. [2007]. For this experiment, the optimal window lengths were determined on the data prior to 1983 to use completely independent data for the model choice and validation. Our method shows less skill for one and two year lead times compared to the assimilated forecast system DEPRESYS from Smith et al. [2007]. For longer lead times the RMSE compares well with that of their DEPRESYS system, and performs significantly better, according to their 90% confidence interval, than their reference forecast, NOASSIM. The reduced skill of our 1-2 year forecasts may be due to the fact that the Smith et al. [2007] model has skill in predicting El Nino, and that it uses a persistence of the sulphate forcing and therefore includes parts of the volcanic forcing. As we only use the “non-volcanic” ensemble for the validation, the eruption of El Chichón in 1982 and Pinatubo in 1991 will decrease our hindcast skill in comparison to theirs.

Smith et al. [2007] further gives the RMSE derived from hindcast experiments on different time averages of the global mean temperature, averaged over all lead times. We perform the same hindcasting experiments, again on the same years used by Smith et al. [2007]. Our RMSE results are 0.106 (IENS) compared to 0.105 (DEPRESYS) for annual averages, 0.059 (IENS) compared to 0.066 (DEPRESYS) for 5-year means and 0.044 (IENS) compared to 0.046 (DEPRESYS) for 9-year means. By construction, the only multi-decadal variability that our model predicts is due to persistence. Since the IENS method performs similar to the model of Smith et al. [2007], which models natural variability for lead times larger than two years, suggests that most of the skill of the DEPRESYS model comes from their assimilated initial conditions.

It should be noted that it is difficult to make such a comparison using only the time period after 1982. As the global mean temperature was dominated by a relatively linear trend in these years this period might be too short to represent the effect of decadal to multidecadal natural variability on the hindcast.

The actual IENS forecast for 2007-2011 is shown in Figure 1 a-c and in Table 1. Compared to the recent decade, GL is predicted to increase more than the other temperature predictions. This is due to the model ensemble mean prediction of a stronger temperature increase in GL than in NH. One reason for this may be a slight decrease in the Atlantic Thermohaline Circulation (THC) in the models as a response to increasing CO_2 [Schmittner et al., 2005]. The THC reduction has a stronger effect on NH than on GL (Knight, Allan et al. 2005) and would therefore partly offset the warming trend in the NH timeseries. For the MDR SST, our model predicts a slight cooling compared to the last five year mean. The reason for this is that the last four years were exceptionally warm compared to the optimal calibration timescale of seven years, and that the amplitude of the externally forced trend in this region is smaller than that of the GL or NH temperature trends. For this reason the RMSE of this forecast, which are given in Table 1, show that the uncertainty of the MDR forecast is high compared to the errors of the other predictions.

.

6. Conclusions

Our simple technique of using the CMIP3 ensemble mean, bias-corrected to the current climate as a prediction for future temperatures, compares favorably with both statistical predictions and the predictions from a complex forecast model by Smith et al. [2007]. We attribute this skill to the combination of a bias-correction, which accounts for the longer-scale natural variability, and the mean of the CMIP3 ensemble, which, while averaging out the internal variability of each model, predicts the response due to anthropogenic forcing. As our technique uses the predictability of the response to anthropogenic forcing it has an advantage predicting variables where anthropogenic effects dominate natural variability.

The results of our quinquennial forecasts, for the global and northern hemispheric mean temperatures of 2007-2011, predict unprecedented warmth. However, a slight decrease in MDR SSTs compared to the last five years is also predicted. Compared to the last decade the global mean temperature is predicted to increase faster than the NH mean temperature which may be due to a slight decrease in the thermohaline circulation which some models are simulating as a response to increasing CO_2.

Since we envision that dynamical forecasting using assimilated initial conditions is actually the future for predictions on these time scales and yet acknowledge the huge technical and computing resources that this requires, we suggest that the presented simple forecasting method can serve as a benchmark for future prediction schemes.

Acknowledgements

We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy We would like to thank Dáithí Stone, the anonymous reviewer and Gerrit Lohmann for helpful comments.

References

Emanuel, K. (2005). "Increasing destructiveness of tropical cyclones over the past 30 years." Nature **436**(7051): 686-688.

Goldenberg, S. B., C. W. Landsea, et al. (2001). "The recent increase in Atlantic hurricane activity: Causes and implications." Science **293**(5529): 474-479.

Hansen, J., R. Ruedy, et al. (2006). "NASA GISS Surface Temperature (GISTEMP) Analysis." trends: A compendium of data on global change.Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tenn., USA.

Knight, J. R., R. J. Allan, et al. (2005). "A signature of persistent natural thermohaline circulation cycles in observed climate." Geophysical Research Letters **32**(20): -.

Lee, T. C. K., F. W. Zwiers, et al. (2006). "Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing." Journal of Climate **19**(20): 5305-5318.

Palmer, T. N., A. Alessandri, et al. (2004). "Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER)." Bulletin of the American Meteorological Society **85**(6): 853-+.

Rayner, N. A., D. E. Parker, et al. (2003). "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century." Journal of Geophysical Research-Atmospheres **108**(D14): -.

Santer, B. D., T. M. L. Wigley, et al. (2006). "Forced and unforced ocean temperature changes in Atlantic and Pacific tropical cyclogenesis regions." Proceedings of the National Academy of Sciences of the United States of America **103**(38): 13905-13910.

Schmittner, A., M. Latif, et al. (2005). "Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations." Geophysical Research Letters **32**(23): -.

Smith, D. M., S. Cusack, et al. (2007). "Improved surface temperature prediction for the coming decade from a global climate model." Science **317**(5839): 796-799.

Solomon, S. e. a. (2007). Climate Change 2007: The Physical Science Basis: Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York,NY,USA, Cambridge University Press.

Zwiers, F. W. (2002). "Climate change - The 20-year forecast." Nature 416(6882): 690-691.

Figure captions

Figure 1. 1a-c show the observed timeseries (thin line), the 5 year mean of the observed timeseries (thick line) and the ensemble mean of the non-volcanic model runs (dashed line). The corresponding indices are MDR SST (a), the NH temperature (b) and GL temperature (c). All timeseries are anomalies from 1951-1980 and the ensemble mean timeseries is shifted by 0.75K for easier visual comparison with the observations. Additionally the 2007-2011 forecast of the IENS method is shown as horizontal thick line. The RMSE associated with each prediction using IENS (white), FLAT (gray) and PERSISTENCE (black) is shown in 1d. The error bar on the IENS RMSE value is the 90% bootstrap confidence interval.

Figure 2. Impact of the bias correction window length on the hindcast skill. The RMSE of the GL temperature, the NH temperature and the MDR SST five year means are shown for the IENS method (continuous line) and for the FLAT method (dashed line)

Figure 3. Dependence of the hindcast skill on lead time. RMSE for annual global mean temperature are shown. a) using IENS-forecast (continuous), FLAT forecast (long dashed) and using PERSISTENCE forecast (dotted). b) as in a) but the validation years are restricted to 1982-2004 to allow for a direct comparison with Figure 1a of Smith et al. [2007].

Tables

Table 1. The predictions for the 2007-2011 surface temperature mean from the IENS technique. Additionally the estimated RMSE of the forecast and the optimal calibration window length used are given.

	GL SAT	NH SAT	MDR SST
Forecast, relative to 1951-1980 (˚C)	0.63	0.77	0.44
Forecast error, RMSE (˚C)	0.084	0.107	0.171
Calibration window length (years)	7	7	7

Figures

Figure 1:

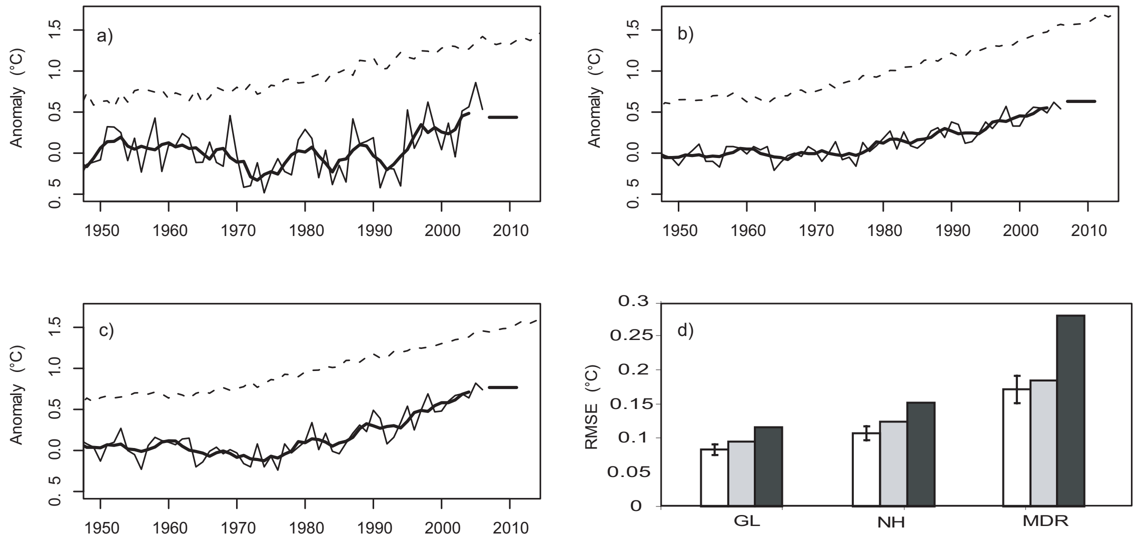


Figure 2:

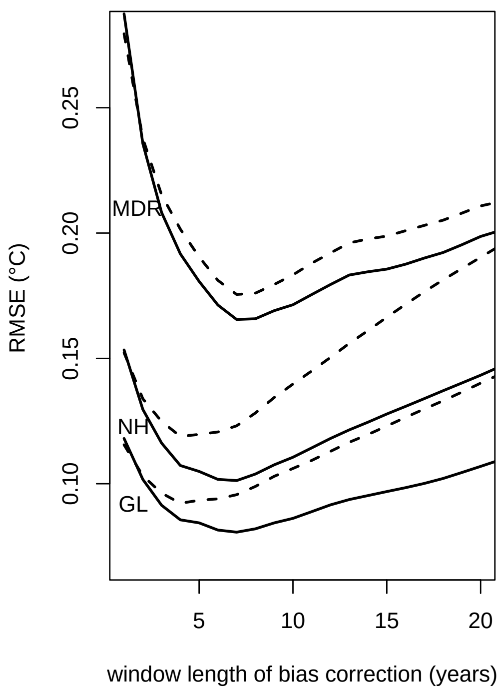
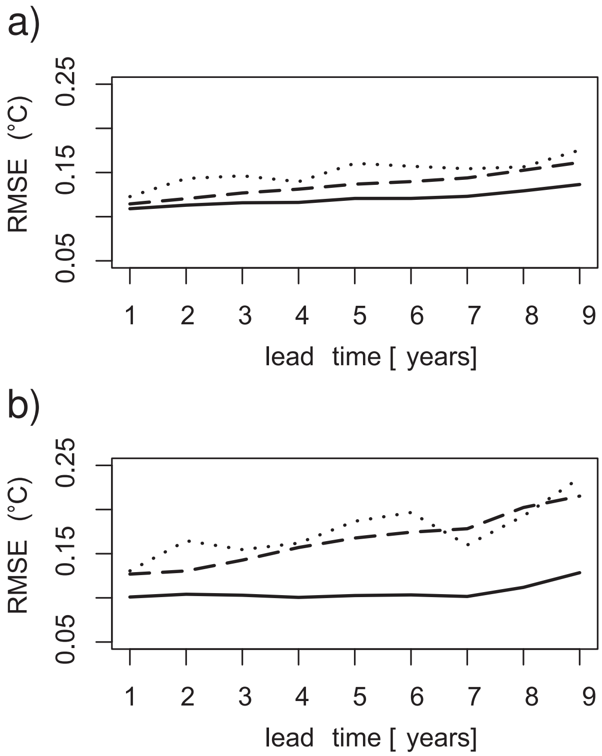

0.25
0.20
0.15
0.10
RMSE (°C)
MDR
NH
GL
5
10
15
20
window length of bias correction (years)

Figure 3:


a)
RMSE (°C)
0.05
0.15
0.25
1
2
3
4
5
6
7
8
9
lead time [years]
b)
RMSE (°C)
0.05
0.15
0.25
1
2
3
4
5
6
7
8
9
lead time [years]

Supporting Online Material for

Interannual Temperature Predictions using the CMIP3 multi-model ensemble mean

Thomas Laepple[1], Stephen Jewson[2], Katie Coughlin[2]

[1]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

[2]Risk Management Solutions, London

E-mail: thomas.laepple@awi.de

Comparison to Smith et al. 2007, using a 16 member ensemble.

As shown in main text, the presented IENS method performs significantly better than the NOASSIM reference approach from Smith et al. 2007 and is comparable to his DEPRESYS approach. Here we investigate whether the skill from IENS is due to the larger ensemble mean, which reduces the remaining natural variability, or due to the bias correction. The full IENS method uses 21 multimodel ensemble members for the years preceding 2000 and 54 ensemble members from 2000 onwards as we use three scenarios for the simulations after 2000.

To test the influence of the ensemble size we investigate a reduced version of IENS by using 16 member ensemble means. The NOASSIM method from Smith et al. [2007] uses 4 ensemble members starting in 4 seasons for each year. As the evaluation is on annual and multiannual timescales, we treat the seasons as ensemble members, and therefore use 16 annual members. For this experiment we restrict ourselves to the SRES A1B scenario. As an exhausting permutation of 16 runs from the available 21 runs is not possible given our current computing power, we calculate the skill for 500 randomly sampled 16 ensemble means.

The results are shown in Figure 1S and 2S. Figure 1S corresponds to Figure 3 of the main manuscript and shows the dependence of the hindcast skill on lead time evaluated on the annual global mean temperature. The effect of the reduced ensemble members is very small and the full IENS result is close to the average of the reduced ensemble experiments. The spread of the results shows the dependence on individual model runs. For lead times larger than two years, every tested combination of model runs has a smaller RMSE than the NOASSIM method from Smith et al. [2007]

In Figure 2S, histograms of the hindcast RMSE are shown evaluated on the same years and the same temporal averages as Smith et al. [2007]. Even with the reduced ensemble size, the RMSE are smaller than the NOASSIM RMSE for all permutations. The skill of the full ensemble IENS method is inside the center of the reduced member skill distribution.

The study using the 16 ensemble members shows that the main skill difference between IENS and NOASSIM from Smith et al. [2007] is caused by the bias correction and not by the larger ensemble size. However one has to note that we are using multimodel ensemble means which could have a positive effect on the hindcast skill compared to single model ensemble means.

References:

Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris, and J.M. Murphy, Improved surface temperature prediction for the coming decade from a global climate model, Science, 317 (5839), 796-799, 2007.

Figure Captions:

Figure 1S Dependence of the hindcast skill on lead time (see Fig. 3 of the main manuscript). RMSE for annual global mean temperature are shown. a) using IENS-forecast (continuous), FLAT forecast (long dashed) and using PERSISTENCE forecast (dotted). b) as in a) but the validation years are restricted to 1982-2004 to allow for a direct comparison with Figure 1a of Smith et al. [2007]. Additionally the results for the 16 ensemble experiments are shown as grey lines.

Figure 2S Histogram of the hindcast skill (RMSE) for the 16 member IENS experiments. The results are shown mean global temperature for annual (a), 5 yr means (b) and 9 yr means (c), averaged over all lead times. The continuous vertical line shows the NOASSIM skill from Smith et al. [2007], the dashed vertical line shows the skill of DEPRESYS from Smith et al. [2007] and the dotted vertical line the skill of the full member IENS hindcast.

Figures:

Figure 1S:

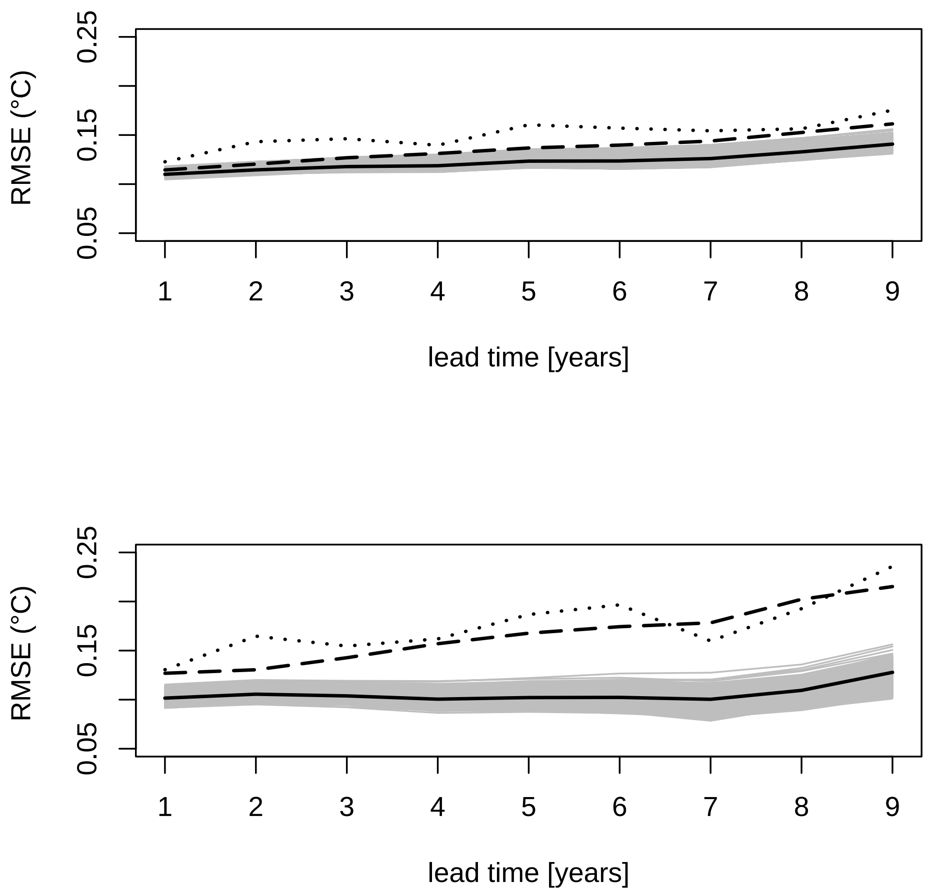
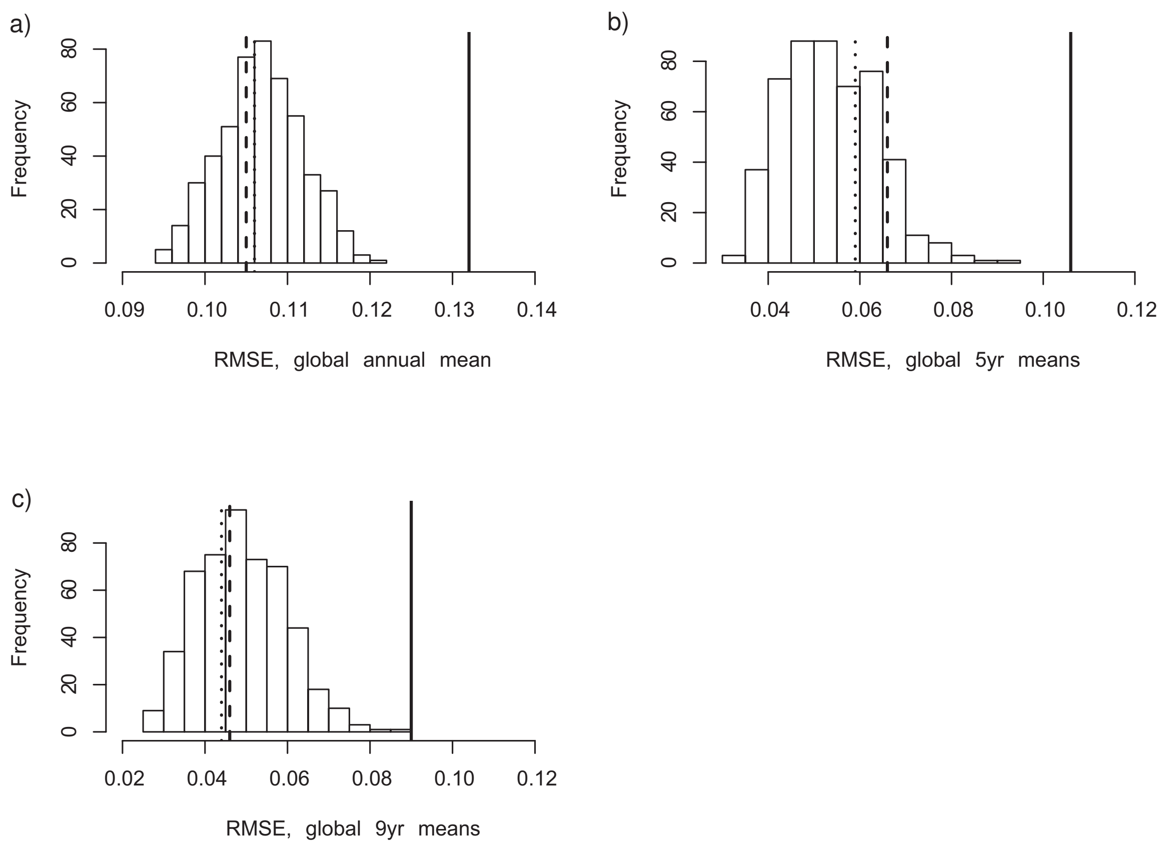


Figure 2S:


a)
Frequency
0
20
40
60
80
0.09
0.10
0.11
0.12
0.13
0.14
RMSE, global annual mean
b)
Frequency
0
20
40
60
80
0.04
0.06
0.08
0.10
0.12
RMSE, global 5yr means
c)
Frequency
0
20
40
60
80
0.02
0.04
0.06
0.08
0.10
0.12
RMSE, global 9yr means