1 **Interannual Temperature Predictions using the CMIP3 multi-model ensemble mean**

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

3 [1]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

4 [2]Risk Management Solutions, London

5 thomas.laepple@awi.de

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24   We present a simple method to make multi-year surface temperature forecasts using the

25   climate change simulations of the CMIP3 database prepared for the IPCC AR4 report. By

26   calibrating the multi-model ensemble mean with current observations, we are able to

27   make skillful interannual forecasts of mean temperatures. The method is validated using

28   extensive hindcast experiments and is shown to perform favorably compared to a recent

29   forecast method based on a global circulation model with assimilated initial conditions.

30   Five year forecasts for the global mean temperature, the Northern Hemispheric mean

31   temperature and the summer sea surface temperatures (SSTs) in the main development

32   region for hurricanes (MDR) are presented.

33

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.

56 We present a very simple approach for interannual temperature forecasts using the

57 existing output from the large ensemble of coupled ocean-atmosphere models which

58 participated in the Coupled Model Intercomparison Project (CMIP3). By calibrating the

59 model output with observed data, we use both the skill of the complex models in

60 forecasting the anthropogenic contribution to changing temperatures and the skill of

61 persistence, which is inherent in the temperature timeseries. Using this precompiled

62 source of information, with appropriate bias corrections, we are able to make skillful

63 interannual temperature predictions and we suggest that this simple prediction technique

64 serve as a benchmark for future prediction experiments.

65

66 We demonstrate our prediction technique on three temperature indices: the annual global

67 mean surface temperature (SAT) which exhibits very small interannual variability due to

68 the large area mean, the Northern Hemispheric mean SAT, and the summer sea surface

69 temperature (SST) in the main development region (MDR). SSTs in this Atlantic region

70 exhibit very strong interannual to multi-decadal variability and are of special interest due

71 to the possible connection to hurricane frequency and intensity (e.g. Goldenberg, Landsea

72 et al. 2001; Emanuel 2005). Forecasts of these indices are given for a five-year outlook

73 and the skill of the interannual forecasts is compared to Smith et al. [2007].

74

75

76 ## *2. Data*

77 We use the annual mean Land-Ocean Temperature anomaly Index for the Northern

78 Hemispheric (NH) and Global mean Temperature (GL) provided by NASA GISS

79   (Hansen, Ruedy et al. 2006) (available at http://data.giss.nasa.gov/gistemp/). The

80   HADISST dataset (Rayner, Parker et al. 2003) is used to extract the MDR SST index (15-

81   70W,10-20N, JAS mean).   We use an anomaly relative to 1951-1980. The three

82   timeseries are shown in Figure 1a-c.

83

84   The model data consists of gridded global monthly SAT and SST from the World

85   Climate Research Programme's Coupled Model Intercomparison Project multi-model

86   dataset (CMIP3) (available at http://www-pcmdi.llnl.gov).

87   We extract mean temperatures over the seasons and regions which correspond to the

88   observational data described above to create analogous time series for each model run.

89   The historical scenario 20C3M as well as the future IPCC-scenarios SRESA1B, SRESA2

90   and SRESB1 are used.

91   ## *3. Forecast method*

92

93   We divide the models into a set which includes historical volcanic forcing and a set

94   without volcanic forcing. As the volcanic forcing has a strong impact on the temperature

95   timeseries, especially on the MDR SST   (Santer, Wigley et al. 2006), but is not

96   predictable in the future, we only use the non-volcanic models in this study to allow for

97   fair hindcasts. The historical 20C3M simulations are merged with the future simulations.

98   The concatenated simulations are then treated as continuous timeseries for the rest of the

99   study. The models BCC-CM1 as well as the SRESB1, CSIRO-Mk3.0 runs were removed

100   from the set to avoid discontinuities in 2000 as they did not restart from the last year of

101   the 20C3M run.

102

103   For the next decade, the differences in the forcing of the scenarios are small (Zwiers

104   2002) so, in order to increase the size of our ensemble, we have included runs from all

105   three. By taking the mean over all the ensemble members of the models and over these

106   three scenarios we are able to remove most of the internal variability of the models. The

107   resulting non-volcanic ensemble mean timeseries are shown in Figure 1a-c together with

108   the observed timeseries.

109

110   In order to create a prediction of a temperature timeseries for the years n+1 onwards, a

111   bias correction is needed to shift the ensemble mean to the current state of the observed

112   temperatures. The current state is estimated using a number of years, N, before the

113   current date, n. The correction then involves subtracting an average of the ensemble

114   mean values over these years (n-N,n-N+1, …, n) and adding an average of the observed

115   values over these years.

116

117   Applying this bias correction, we predict future temperature values from simulated values

118   for the years n+1 onwards. We call this IPCC/CMIP3 ensemble based method IENS. The

119   IENS approach is similar to the reference method NOASSIM from Smith et al. [2007],

120   but the use of our optimized N year baseline takes into account slow natural variability.

121

122   As reference predictions, we provide an optimal persistence forecast which is the mean of

123   the N years before the current date (we call this FLAT), and a simple persistence estimate

124   which is the value of the year before the forecast (we call this PERSISTENCE). By

125   construction of the FLAT forecast, the IENS forecast will have higher skill if, on average,

126   the trend of the ensemble mean is realistic. We note here that a linear trend prediction,

127   modeled using an optimized window length for the trend fit, was initially included for

128   comparison. Although not shown here, the skill of this forecast was always less than that

129   of the IENS method, and often less than for the FLAT method.

130

131   Obviously the forecasts IENS and FLAT depend on the calibration window length, N.

132   The optimal N depends on the properties of the timeseries as well as on the lead time and

133   is determined by hindcasting on the historical data where N is defined to be the number

134   of years which minimizes the root mean squared error (RMSE). Figure 2 shows the

135   dependence of the RMSE of 5-year mean hindcasts on N based on hindcasts from 1930-

136   2006.

137

138   In terms of forecast error, there is an optimal calibration window, which in this case is

139   seven years, for all of the IENS hindcasts. The RMSE of the IENS methods are lower

140   than the RMSE of the FLAT method which shows that the CMIP3 ensemble mean adds

141   skill to the forecast. How can we explain the shape of the calibration window length

142   dependence? For very short calibration windows, the mean state is not well estimated and

143   its large variance dominates the RMSE. Therefore, as the calibration window increases

144   the RMSE decreases approximately as the standard error of the mean decreases

145    $(1/\sqrt{N})$. For long calibration periods, biases between the observations and the model

146   mean, due to natural variability or structural errors of the models, become important and

147   contribute to an increasing RMSE. A balance between these effects gives the minima

148   seen in Figure 2.

149

150   *4. Validation method*

151

152   To compare prediction methods, we use the RMSE of hindcast experiments. For each

153   hindcast, the window length for the FLAT and the IENS method are re-estimated using

154   all data except an interval of 10 years surrounding the years to be hindcast. This is done

155   to minimize the artificial inflation of forecast skill which occurs when the window length,

156   N, is estimated using the same data as is used to validate the forecast.

157   Because there is a limited hindcast period, we also supply the 90% bootstrapping

158   confidence intervals to estimate the uncertainty of the RMSE. These confidence intervals

159   are derived by randomly sampling (with replacement) m hindcast errors where m is the

160   total number of hindcasts (e.g. for the quinquennial forecast, m = 73). This is repeated

161   10,000 times and a RMSE is estimated each time to derive a distribution.

162

163   This will not, however, account for systematic errors that may be found in our estimate of

164   the prediction error. There are some reasons why the hindcast RMSE may be a

165   conservative estimate of the forecast RMSE:

166   1.) If there are no volcanoes during the forecast period, the error may be smaller than

167   estimated since the hindcast is performed over past periods which did include volcanoes.

168 2.) The mean of three scenarios is used for the forecast, but there is only one scenario for

169 most hindcast years. Therefore, the residual of the internal variability is smaller for years

170 after 2000, which might reduce the forecast error.

171 3.) We perform the validation on all available years (1930-2006) to represent the natural

172 variability. However, one can argue that the higher ratio, of externally forced change to

173 natural variability, in recent years will reduce the future error of the IENS approach.

174

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

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

177 years of the hindcast experiment.

178   2.) Some of the model results may be tuned to the observational period causing the IENS

179 hindcasts to be closer to the observations and artificially lowering the RMSE.

180

5. Results of validation and forecasts

182

183 The estimated RMSE for the different methods are compared in Figure 1d. These

184 RMSEs are slightly higher than the minimum RMSE in Figure 1a-c since these errors

185 also include the uncertainty in the window length estimation. Figure 1d shows that the

186 IENS forecast is generally more accurate than the reference methods, FLAT and

187 PERSISTENCE. However, the 90% bootstrap confidence interval shown by the error

188 bars on the IENS value indicates that the ensemble mean forecast is significantly better

189 than the PERSISTENCE forecast but not necessarily better than the FLAT forecast for

190 the MDR SSTs. This is understandable if much of our prediction skill comes from the

191   bias-correction, or estimate of the current state. The added skill due to the anthropogenic

192   changes modeled by the ensemble mean is most obvious in the global mean and NH

193   mean temperature where natural variability is small due to the larger spatial averaging.

194   This result is consistent with the results from (Lee, Zwiers et al. 2006), who found

195   decadal climate prediction skill of the global mean temperature due to changes in

196   anthropogenic forcing.

197

198   Next we compare the skill of our method and the method of Smith et al. (2007). They use

199   the HadCM3 model, with assimilated inital conditions, to predict temperatures out to 9

200   years.

201   Figure 3 shows the RMSE of annual mean global temperature forecasts using the IENS,

202   FLAT and PERSISTENCE method for lead times from 1-9 years. The hindcasts are

203   based on the period 1939-2006. 1939 is chosen as the initial year for the hindcasts as we

204   test window lengths up to 30 years. The IENS method shows the most skill for all lead

205   times and all three forecast methods show a decrease in skill for longer lead times. The

206   difference between FLAT and PERSISTENCE RMSE decreases with lead time whereas

207   the difference between IENS and FLAT increases with lead time. The reason for this is

208   that when the bias dominates, for the FLAT and PERSISTENCE models, the better

209   estimate of the mean state becomes less important.

210   Since IENS predicts a realistic trend on average, the increase in RMSE with lead time is

211   slower. In Figure 3b we show the same results using the hindcast years 1983-2004, as in

212   Smith et al. [2007]. It can therefore be directly compared to Figure 1a) of Smith et al.

213   [2007]. For this experiment, the optimal window lengths were determined on the data

214   prior to 1983 to use completely independent data for the model choice and validation.

215   Our method shows less skill for one and two year lead times compared to the assimilated

216   forecast system DEPRESYS from Smith et al. [2007]. For longer lead times the RMSE

217   compares well with that of their DEPRESYS system, and performs significantly better,

218   according to their 90% confidence interval, than their reference forecast, NOASSIM. The

219   reduced skill of our 1-2 year forecasts may be due to the fact that the Smith et al. [2007]

220   model has skill in predicting El Nino, and that it uses a persistence of the sulphate forcing

221   and therefore includes parts of the volcanic forcing. As we only use the "non-volcanic"

222   ensemble for the validation, the eruption of El Chichón in 1982 and Pinatubo in 1991 will

223   decrease our hindcast skill in comparison to theirs.

224

225   Smith et al. [2007] further gives the RMSE derived from hindcast experiments on

226   different time averages of the global mean temperature, averaged over all lead times. We

227   perform the same hindcasting experiments, again on the same years used by Smith et al.

228   [2007]. Our RMSE results are 0.106 (IENS) compared to 0.105 (DEPRESYS) for annual

229   averages, 0.059 (IENS) compared to 0.066 (DEPRESYS) for 5-year means and 0.044

230   (IENS) compared to 0.046 (DEPRESYS) for 9-year means. By construction, the only

231   multi-decadal variability that our model predicts is due to persistence. Since the IENS

232   method performs similar to the model of Smith et al. [2007], which models natural

233   variability for lead times larger than two years, suggests that most of the skill of the

234   DEPRESYS model comes from their assimilated initial conditions.

235

236   It should be noted that it is difficult to make such a comparison using only the time

237   period after 1982. As the global mean temperature was dominated by a relatively linear

238   trend in these years this period might be too short to represent the effect of decadal to

239   multidecadal natural variability on the hindcast.

240   The actual IENS forecast for 2007-2011 is shown in Figure 1 a-c and in Table 1.

241   Compared to the recent decade, GL is predicted to increase more than the other

242   temperature predictions. This is due to the model ensemble mean prediction of a stronger

243   temperature increase in GL than in NH. One reason for this may be a slight decrease in

244   the Atlantic Thermohaline Circulation (THC) in the models as a response to increasing

245   CO_2 [Schmittner et al., 2005]. The THC reduction has a stronger effect on NH than on

246   GL (Knight, Allan et al. 2005) and would therefore partly offset the warming trend in the

247   NH timeseries. For the MDR SST, our model predicts a slight cooling compared to the

248   last five year mean. The reason for this is that the last four years were exceptionally

249   warm compared to the optimal calibration timescale of seven years, and that the

250   amplitude of the externally forced trend in this region is smaller than that of the GL or

251   NH temperature trends. For this reason the RMSE of this forecast, which are given in

252   Table 1, show that the uncertainty of the MDR forecast is high compared to the errors of

253   the other predictions.

254   .

255

256

257

6. Conclusions

258   *6. Conclusions*

259

260

261   Our simple technique of using the CMIP3 ensemble mean, bias-corrected to the current

262   climate as a prediction for future temperatures, compares favorably with both statistical

263   predictions and the predictions from a complex forecast model by Smith et al. [2007].

264   We attribute this skill to the combination of a bias-correction, which accounts for the

265   longer-scale natural variability, and the mean of the CMIP3 ensemble, which, while

266   averaging out the internal variability of each model, predicts the response due to

267   anthropogenic forcing. As our technique uses the predictability of the response to

268   anthropogenic forcing it has an advantage predicting variables where anthropogenic

269   effects dominate natural variability.

270

271   The results of our quinquennial forecasts, for the global and northern hemispheric mean

272   temperatures of 2007-2011, predict unprecedented warmth. However, a slight decrease

273   in MDR SSTs compared to the last five years is also predicted. Compared to the last

274   decade the global mean temperature is predicted to increase faster than the NH mean

275   temperature which may be due to a slight decrease in the thermohaline circulation which

276   some models are simulating as a response to increasing CO_2.

277

278   Since we envision that dynamical forecasting using assimilated initial conditions is

279   actually the future for predictions on these time scales and yet acknowledge the huge

280   technical and computing resources that this requires, we suggest that the presented simple

281   forecasting method can serve as a benchmark for future prediction schemes.

282

283

284

285

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

References

Emanuel, K. (2005). "Increasing destructiveness of tropical cyclones over the past 30

years." <u>Nature</u> **436**(7051): 686-688.

Goldenberg, S. B., C. W. Landsea, et al. (2001). "The recent increase in Atlantic

hurricane activity: Causes and implications." <u>Science</u> **293**(5529): 474-479.

Hansen, J., R. Ruedy, et al. (2006). "NASA GISS Surface Temperature (GISTEMP)

Analysis." <u>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</u>.

Knight, J. R., R. J. Allan, et al. (2005). "A signature of persistent natural thermohaline

circulation cycles in observed climate." <u>Geophysical Research Letters</u> **32**(20): -.

Lee, T. C. K., F. W. Zwiers, et al. (2006). "Evidence of decadal climate prediction skill

resulting from changes in anthropogenic forcing." <u>Journal of Climate</u> **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)." <u>Bulletin of

the American Meteorological Society</u> **85**(6): 853-+.

326 Rayner, N. A., D. E. Parker, et al. (2003). "Global analyses of sea surface temperature,

327   sea ice, and night marine air temperature since the late nineteenth century."

328   Journal of Geophysical Research-Atmospheres **108**(D14): -.

329 Santer, B. D., T. M. L. Wigley, et al. (2006). "Forced and unforced ocean temperature

330   changes in Atlantic and Pacific tropical cyclogenesis regions." Proceedings of the

331   National Academy of Sciences of the United States of America **103**(38): 13905-

332   13910.

333 Schmittner, A., M. Latif, et al. (2005). "Model projections of the North Atlantic

334   thermohaline circulation for the 21st century assessed by observations."

335   Geophysical Research Letters **32**(23): -.

336 Smith, D. M., S. Cusack, et al. (2007). "Improved surface temperature prediction for the

337   coming decade from a global climate model." Science **317**(5839): 796-799.

338 Solomon, S. e. a. (2007). Climate Change 2007: The Physical Science Basis: Working

339   Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on

340   Climate Change. Cambridge, United Kingdom and New York,NY,USA,

341   Cambridge University Press.

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

343   691.

344

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
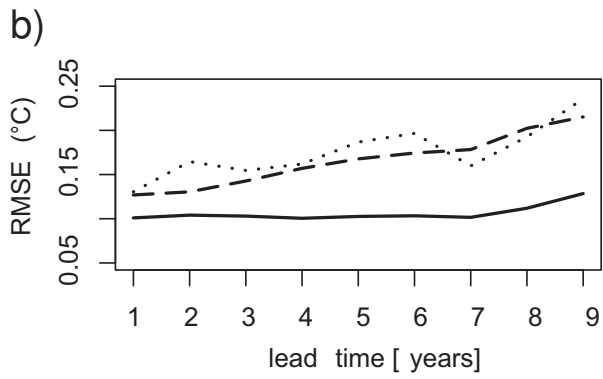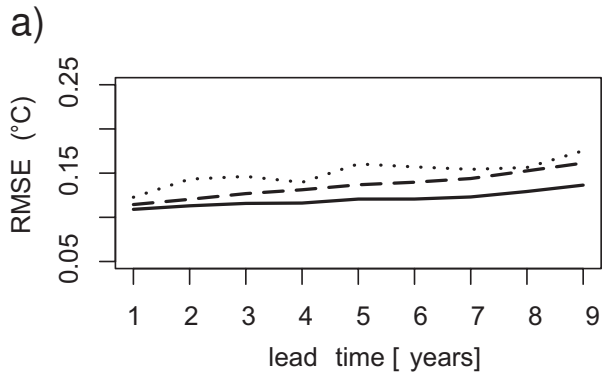
Figure 1:

Figure 2:

Figure 3:

a)

b)
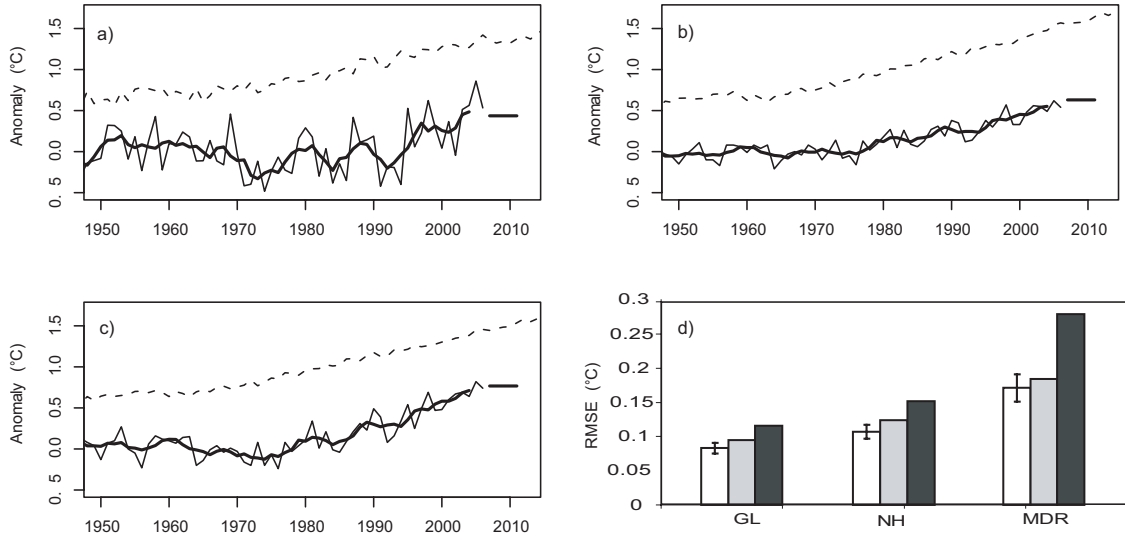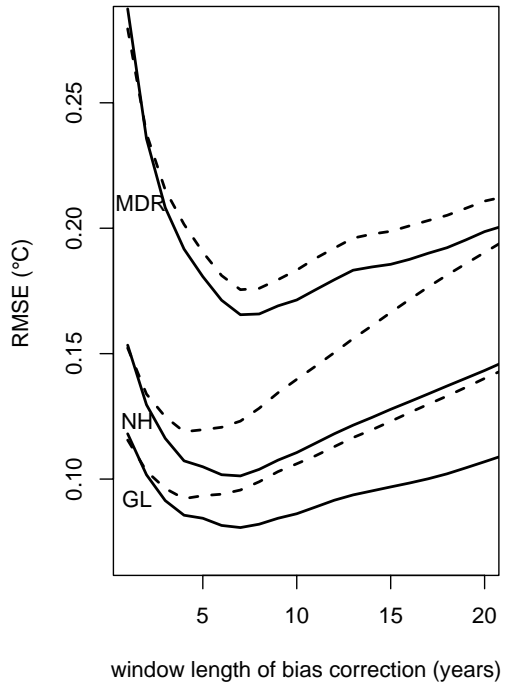
Supporting Online Material for

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.
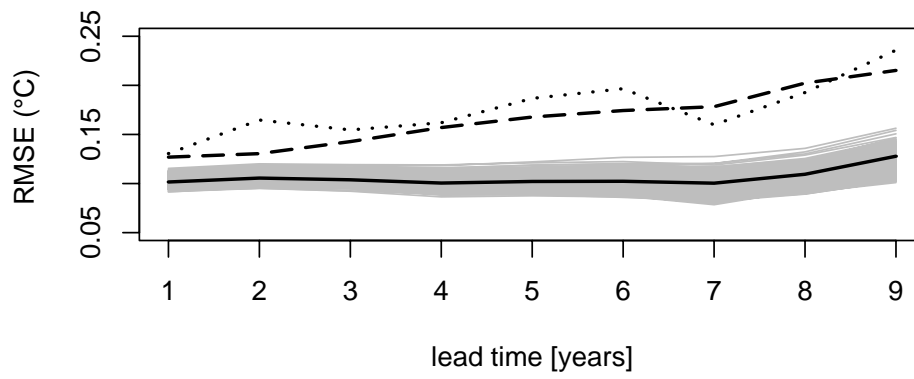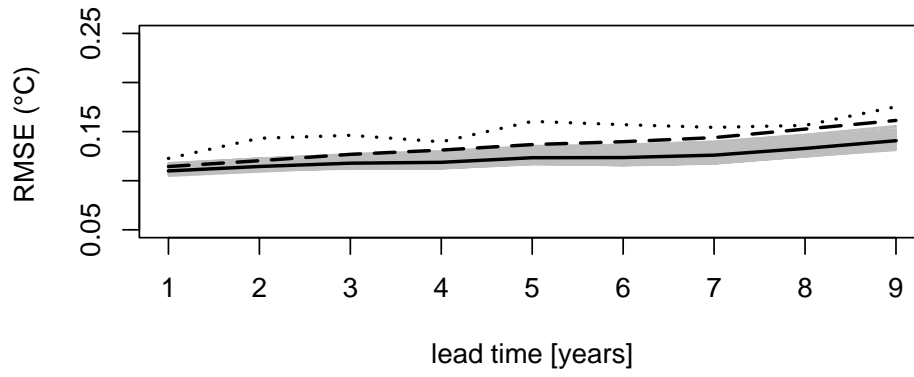
Figures:

Figure 1S:

Figure 2S:

a)

RMSE, global annual mean

b)

RMSE, global 5yr means

c)

RMSE, global 9yr means