An evaluation of decadal probability forecasts from state-of-the-art climate models - Supplementary Material

Emma B. Suckling and Leonard A. Smith

October 21, 2013

1. Introduction

The following material is a supplement to 'An evaluation of decadal probability forecasts from state-of-the-art climate models', in which the perfomance of simulation models from Stream 2 of the ENSEMBLES decadal hindcasts (Doblas-Reyes et al. 2010) are contrasted with the empirical dynamic climatology (DC) model over global and Giorgi region scales. Further details about transforming ensemble simulations into probabilistic distributions are presented below in Section 2. In Section 3 it is shown that the DC empirical model outperforms the ENSEMBLES simulation models by several bits at most lead times and for every region studied. In Section 4 the robustness of the results in the main manuscript are evaluated by using alternative proper scoring rules, namely the proper linear (PL) and continuous ranked probability scores (CRPS). It is shown that the results are robust to the scoring rule chosen. Finally, in Section 5 the performance of alternative empirical models are considered, namely a ‘Prelaunch linear trend’ approach and ‘Prelaunch DC model’. It is shown that the Prelaunch DC model performs to a similar quality as the standard DC approach employed in the main manuscript, and is robust to the kernel parameters and anchor year chosen to fit the model. Further details about generating the probabilistic DC forecasts and the robustness of the results to the model parameter choices are also provided in Section 5.

2. Probabilistic forecast distributions for the ENSEMBLES simulation models

Figures 1, 2 and 3 illustrate the probabilistic forecast distributions for the ENSEMBLES simulation models, generated by kernel dressing the ensemble members as described in the main manuscript and below under cross-validation (the forecast distributions for HadGem2 are illustrated in figure 3 in the main manuscript).

Information contamination is a significant concern in the evaluation of decadal forecasts. Given that the total duration of hindcast experiments is typically fifty years, there are very few independent decadal periods in the forecast-outcome archive. Cross-validation approaches attempt to maximise the size of the forecast-outcome archive (to increase statistical significance) while avoiding the use of information from a given forecast target period being used in the evaluation of that forecast. It is crucial to also avoid information contamination by inadvertently using information from the target decade when interpreting the ensemble into a forecast distribution (Bröcker and Smith 2008). This cannot be done rigorously in the case of simulation models, as the structure and parameters of the models themselves have evolved in light of the observations of the last fifty years. The true-leave-one-out cross-validation procedure described in the main maunuscript avoids any explict use of data from within the target forecast period, even as its implicit use cannot be avoided. In practice this is achieved by leaving out the target decade, then using a standard leave-one-out procedure to fit the kernel parameters for each forecast in turn.

Figure 4 shows an example of the kernel parameters used for the HadGem2 model, fitted using the true-leave-one-out protocol. The top two panels of figure 4 illustrate the mean Ignorance score as a function of kernel width over the full set of hindcast simulations (*i.e.* with no cross-validation) for lead time one and lead time six. The vertical bars indicate the values of the kernel width parameter that were used for each forecast using the true-leave-one-out approach. In both cases the fact that fewer than nine vertical bars are visible indicates that several of the forecasts were generated using the same kernel width values. Note that at lead time six for HadGem2 the kernel width values used are much smaller than for lead time one (and for all other lead times). In this particular case the model is rewarded for a forecast distribution that has kernel widths much smaller than the standard deviation of the ensemble spread.

The bottom panels of figure 4 show the mean Ignorance as a function of kernel offset over the full set of hindcast simulations. Once again the vertical bars indicate the values of offset that were used for the individual forecasts, based on minimising Ignorance through the true-leave-one out protocol. Once again, at lead time six the fitting protocol favours a kernel offset under true-leave-one-out cross-validation that falls outside the minimum Ignorance value without cross-validation. The result for lead time one is typical of the kernel offset values attained for the other lead times.

3. Regional analysis

Figures 5 to 25 show Ignorance as a function of lead time for each of the ENSEMBLES models relative to the DC empirical model for surface air temperature over each of the land-based Giorgi regions (Giorgi 2002). At Giorgi region scales the decadal probability forecasts from the ENSEMBLES models perform to a similar quality as for the global mean temperature in some cases, or significantly worse in others. In some regions and at some lead times DC outperforms the ENSEMBLES models by more than 4 bits; DC placing over 16 (2^4) times more probability mass on the verification than the simulation model. In these figures no simulation model demonstrates skill significantly above the DC model for any lead time or any region; positive values of the relative Ignorance performance measure are reported in all of the cases below.

4. Robustness to the peformance measure

While Ignorance is effectively the only proper local score for the evaluation of probability forecasts (Good 1952), there are a variety of other proper scores that are commonly used in forecast evaluation (Jolliffe and Stephenson 2003). Figures 26 and 27 demonstrate that the results presented in the main text for global mean surface temperature are robust when considered under two alternative scores, the Proper Linear score (PL) and the Continuous

Ranked Probability Score (CRPS) (Jolliffe and Stephenson 2003). In each of these cases, the lower the score the better the forecast. In each case all the models are ranked similarly by the different scores, with DC demonstrating lower scores compared to the ENSEMBLES models.

5. Alternative empirical models

The use of hindcasts in forecast evaluation unavoidably introduces information contamination, as the target of the hindcast is known when the hindcast is made. Thus it is useful to demonstrate that the results of hindcast evaluation are robust to variations in the parameters and even the structure of empirical models, as doing so can identify cases where the hindcast system may have been over-fit in-sample. For the DC empirical model presented in this paper, all data from each target decade being forecast was withheld when constructing that forecast to avoid information contamination. Further avoidance of such information contamination can be achieved in the case of empirical models by using only data from a period *prior* to each forecast launch date and by using a simple model structure. In this section, two Prelaunch empirical models (defined in the main text) are illustrated below, and their robustness to the model parameters examined.

The Prelaunch Dynamic Climatology (Prelaunch DC) model is structurally identical to the DC model of the main manuscript, however only inputs dated before the launch date are used either in the ensemble forecast or in its interpretation into a probability distribution, and so on. While the kernel width used in the standard DC model is determined by cross-validation, this need not be done for the Prelaunch DC model as only the observations available before the forecast launch time are used.

Examining the of the score to variations in the parameters can reveal overfitting. Figure 28 shows the skill of the Prelaunch DC for values of the kernel width ranging from 0.02 to 0.16 for forecast lead times of one to ten years. Ignorance relative to the standard DC model is shown. The sensitivity of the Prelaunch DC model to variation in the starting date for the forecast-outcome archive (not shown) is less than the sensitivity to the kernel width. Start dates from 1900 to 1950 were considered; the later start dates tend to yield more skilful models. The Prelaunch DC discussed in the main text uses a start date of 1950 and a width of 0.08, although this value does not correspond to the lowest in-sample skill - as shown in figure 29. Furthermore the ensemble interpretation of the simulations models reported in this paper use data both before and after the target window, giving those simulation models an unquantified advantage over the empirical models defined here.

Figure 29 shows the mean Ignorance score over the set of DC and Prelaunch DC hindcasts as a function of the kernel width parameter. The panels on the left of figure 29 correspond to lead times one (a), six (c) and ten (e) respectively for the standard DC model, and the panels on the right correspond to the same lead times for the Prelaunch DC model. In each case the vertical bars correspond to the values of kernel spread adopted for each model in the main manuscript (note that for the standard DC model these values were attained under true-leave-one-out cross-validation and for the Prelaunch DC model a value of 0.08 was chosen since cross-validation is not necessary in this case). The fact that there is no significant difference in skill between the standard DC and Prelaunch DC models over a range of kernel dressing parameters indicates that the overall conclusions drawn from the ENSEMBLES model evaluations are not overly sensitive to the particular choice of DC or

Prelaunch DC model parameters.

A Prelaunch trend model is also discussed in the main text. This model is fully defined by the initial time anchor from which the trend is estimated. Figure 30 shows the skill of this model relative to the standard DC model for several anchor times between 1900 and 1950. The results in the main text use the 1950 anchor time. It is shown that although there is some sensitivity to the anchor time, all the Prelaunch trend models are generally less skillful than the standard DC model.

The figures presented in this supplementary material demonstrate that the skill of the empirical models is robust under relatively large variations in their free parameters. This level of skill remains comparable with, and in some cases superior to, that of the simulation models from ENSEMBLES.

REFERENCES

Bröcker, J. and L. A. Smith, 2008: From ensemble forecasts to predictive distributions. *Tellus A*, **60 (4)**, 663–678.

Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Forecast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. *Technical Memorandum ECMWF*, **621**.

Giorgi, F., 2002: Variability and trends of sub-continental scale surface climate in the twentieth century. part i: observations. *Climate Dynamics*, **18**, 675–691.

Good, I. J., 1952: Rational decsions. *Journal of the Royal Statistical Society*, **XIV (1)**, 107–114.

Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast verification: A practitioner's guide in atmospheric science*. John Wiley and Sons Ltd.

List of Figures

1 Forecast distributions for IFS/HOPE (ECMWF) for the 5-95^{th} percentile. The HadCRUT3 observed temperatures are shown in blue. Each forecast is ten years long and they are launched every five years. To avoid overlap of the fan charts they are presented on two panels. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside the 5-95th percentile of the predicted distributions. 16

2 Forecast distributions for ARPEGE/OPA (CERFACS) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside the 5-95th percentile of the predicted distributions. 17

3 Forecast distributions for ECHAM5 (IFM-GEOMAR) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature falls outside the 5-95th percentile of the predicted distributions on several occasions. 18

4 Ignorance as a function of kernel dressing parameters over the full set of hindcast simulations (*i.e.* with no cross-validation) for the HadGem2 model at lead time one (a and c) and lead time six (b and d). The top panels (a and b) show the score as a function of the kernel width parameter and the bottom panels (c and d) show the score as a function of the kernel offset parameter. The vertical bars in each case illustrate the kernel parameters obtained for each individual forecast under true-leave-one-out cross-validation. That there are fewer than nine vertical bars indicates that the kernel parameter values shown were obtained for several forecasts in the set. Results for lead times two to five and seven to ten (not shown) are similar to those shown for lead time one. 19

5 Ignorance of the ENSEMBLES simulation models relative to the DC model for Alaska. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 20

6 Ignorance of the ENSEMBLES simulation models relative to the DC model for Amazon Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 21

7 Ignorance of the ENSEMBLES simulation models relative to the DC model for Australia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 22

8 Ignorance of the ENSEMBLES simulation models relative to the DC model for Central America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 23

9 Ignorance of the ENSEMBLES simulation models relative to the DC model for Central Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 24

10 Ignorance of the ENSEMBLES simulation models relative to the DC model for Central North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 25

11 Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 26

12 Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 27

13 Ignorance of the ENSEMBLES simulation models relative to the DC model for East Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 28

14 Ignorance of the ENSEMBLES simulation models relative to the DC model for Greenland. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 29

15 Ignorance of the ENSEMBLES simulation models relative to the DC model for Mediterranian Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 30

16 Ignorance of the ENSEMBLES simulation models relative to the DC model for North Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 31

17 Ignorance of the ENSEMBLES simulation models relative to the DC model for Northern Europe. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 32

18 Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 33

19 Ignorance of the ENSEMBLES simulation models relative to the DC model for Sahara. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 34

20 Ignorance of the ENSEMBLES simulation models relative to the DC model for South Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 35

21 Ignorance of the ENSEMBLES simulation models relative to the DC model for Southeast Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 36

22 Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern South America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 37

23 Ignorance of the ENSEMBLES simulation models relative to the DC model for Tibet. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 38

24 Ignorance of the ENSEMBLES simulation models relative to the DC model for Western Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 39

25 Ignorance of the ENSEMBLES simulation models relative to the DC model for Western North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models. 40

26 Proper linear score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better foecasts. The DC model is shown to outperform the simulations models at most lead times. 41

27 CRPS score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better forecasts. The DC model is shown to outperform the simulations models at most lead times. 42

28 Ignorance of the Prelaunch DC empirical model with kernel widths as labelled relative to the cross-validation DC model. Increasing the kernel width parameter from 0.02 to 0.16 results in a loss of skill of approximately half a bit, although for the kernel width value used in this paper (0.08) there is shown to be no significant loss of skill relative to the standard DC model. 43

29 Ignorance as a function of the kernel width parameter over the full set of hindcast simulations (*i.e.* with no cross-validation) for the DC (left panels) and Prelaunch DC (right panels) models at lead time one (a and b), six (c and d) and ten (e and f). The vertical bars in each case illustrate the kernel width parameters employed in the main manuscript. In the DC model parameters were attained through true-leave-one-out cross-validation. In the Prelaunch DC model a kernel spread value of 0.08 was chosen for comparison with DC and to test the robustness of the results to choices in the parameters for ensemble interpretation (although this value does not correspond to the lowest value of in-sample skill). 44

30 Ignorance of the Prelaunch trend empirical model for different anchor times relative to the cross-validation DC model. Scores above zero indicate that DC outperforms the Prelaunch Trend model by up to half a bit at early lead times, and up to two bits (DC placing up to 4 times more probability on the observed outcome than the Prelaunch Trend model) up to ten years ahead, depending on the anchor year for the trend model. 45

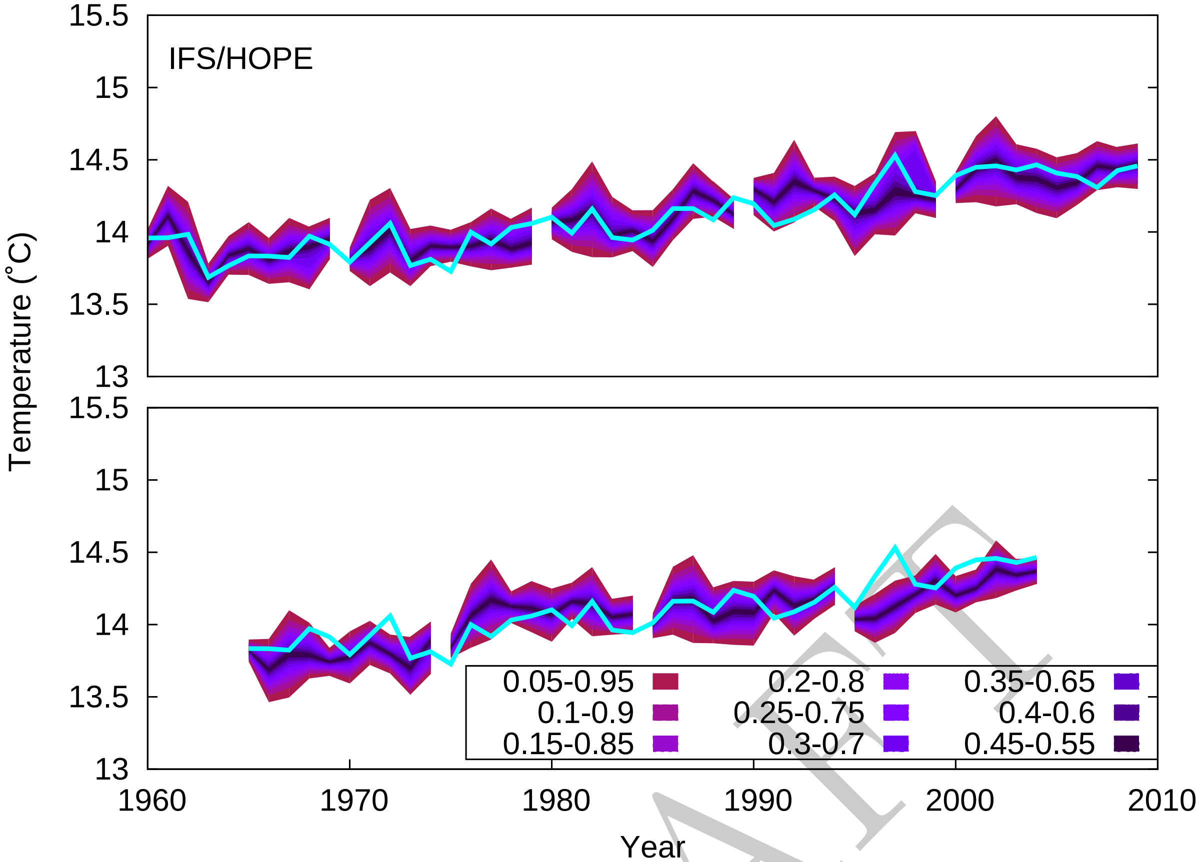

FIG. 1. Forecast distributions for IFS/HOPE (ECMWF) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. Each forecast is ten years long and they are launched every five years. To avoid overlap of the fan charts they are presented on two panels. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside the 5-95th percentile of the predicted distributions.

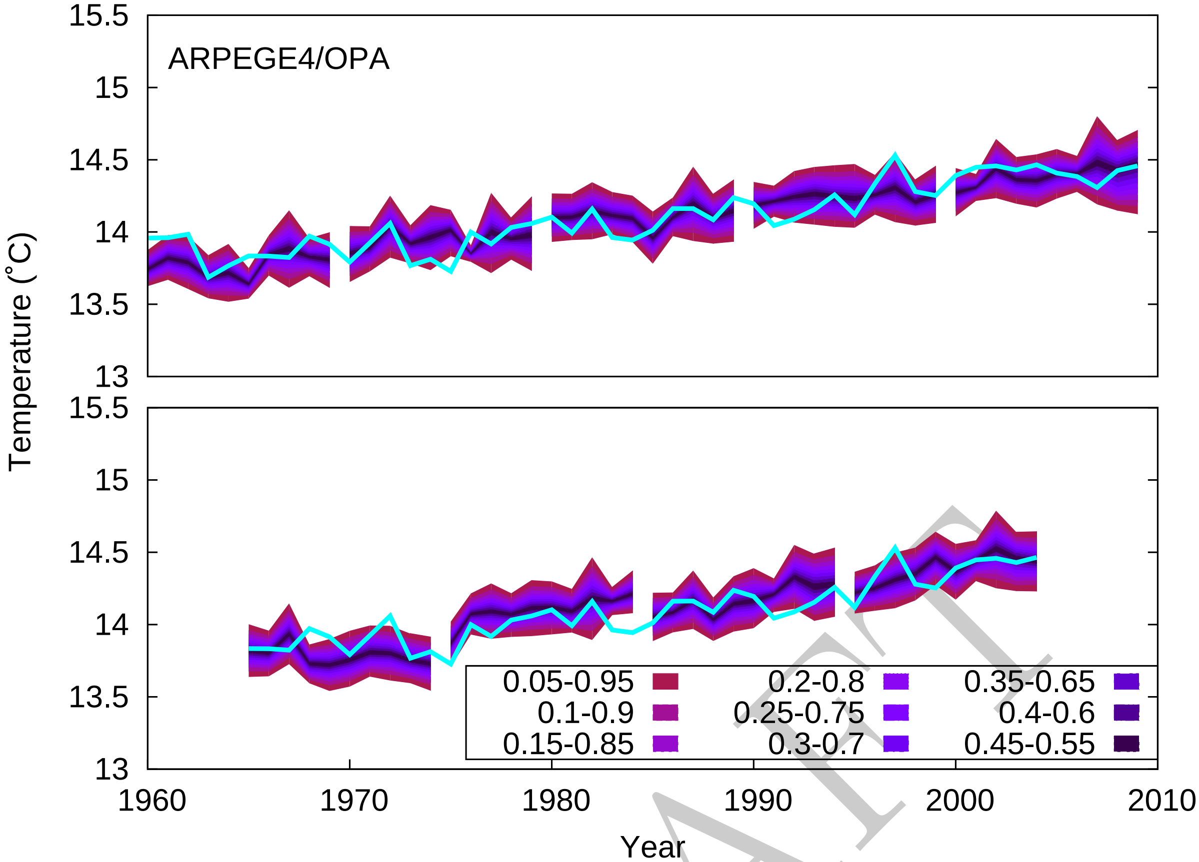

Fig. 2. Forecast distributions for ARPEGE/OPA (CERFACS) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside the 5-95th percentile of the predicted distributions.

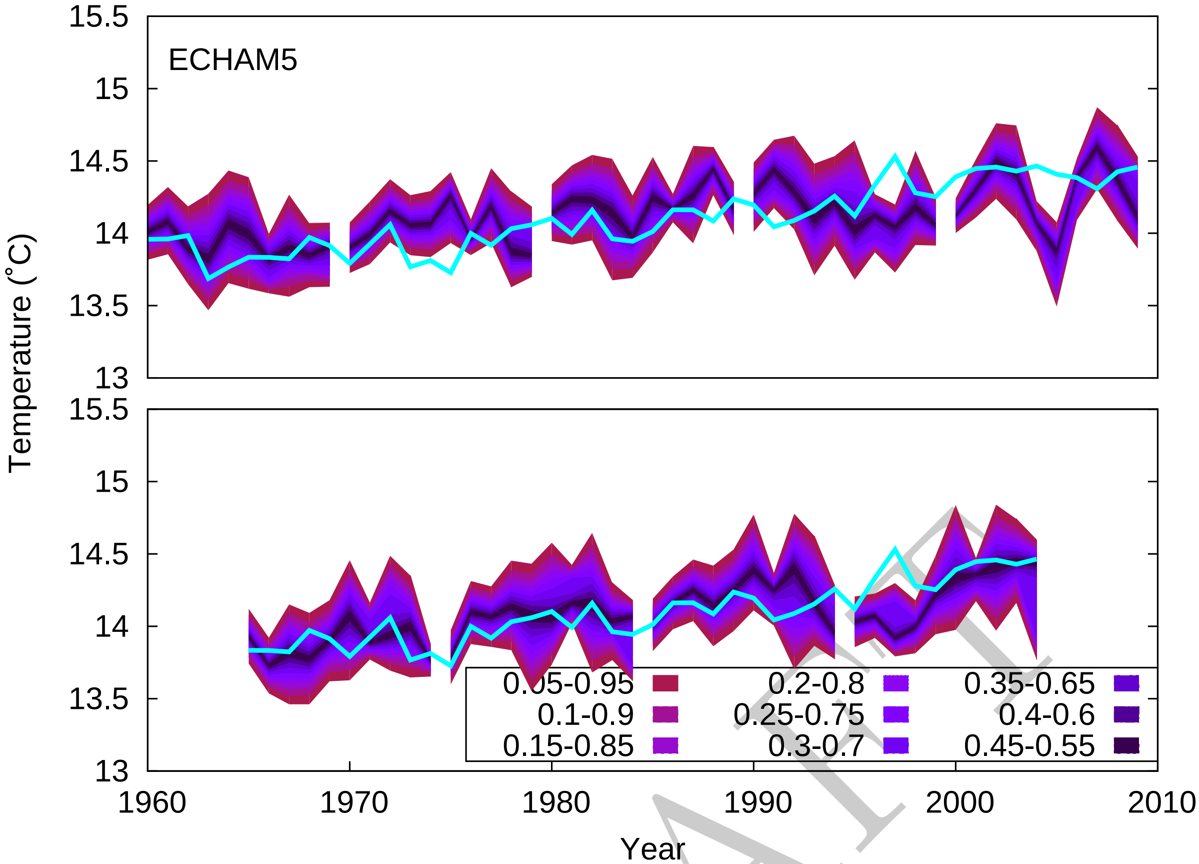

FIG. 3. Forecast distributions for ECHAM5 (IFM-GEOMAR) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature falls outside the 5-95th percentile of the predicted distributions on several occasions.

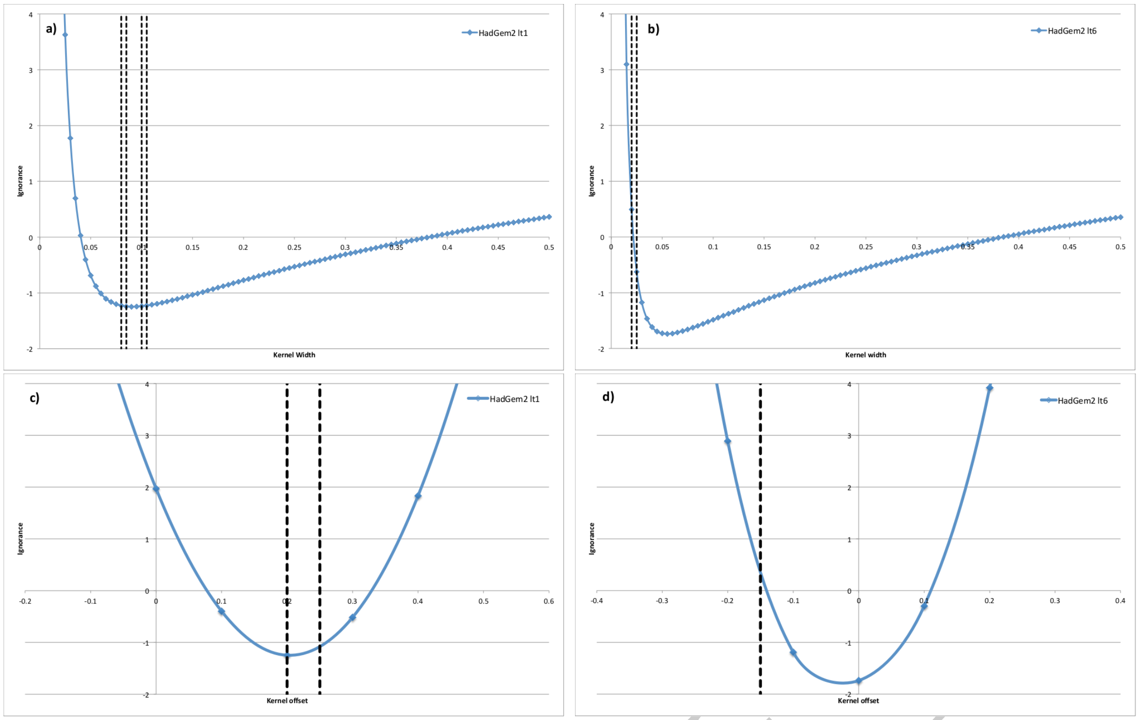

FIG. 4. Ignorance as a function of kernel dressing parameters over the full set of hindcast simulations (*i.e.* with no cross-validation) for the HadGem2 model at lead time one (a and c) and lead time six (b and d). The top panels (a and b) show the score as a function of the kernel width parameter and the bottom panels (c and d) show the score as a function of the kernel offset parameter. The vertical bars in each case illustrate the kernel parameters obtained for each individual forecast under true-leave-one-out cross-validation. That there are fewer than nine vertical bars indicates that the kernel parameter values shown were obtained for several forecasts in the set. Results for lead times two to five and seven to ten (not shown) are similar to those shown for lead time one.

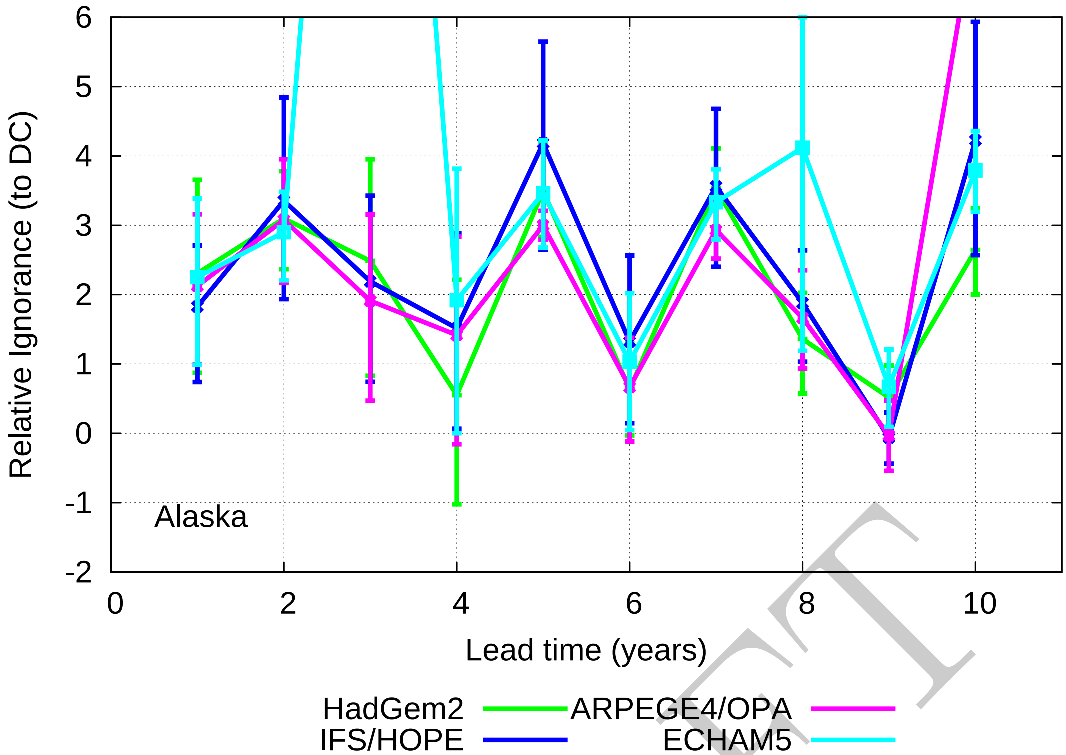

FIG. 5. Ignorance of the ENSEMBLES simulation models relative to the DC model for Alaska. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

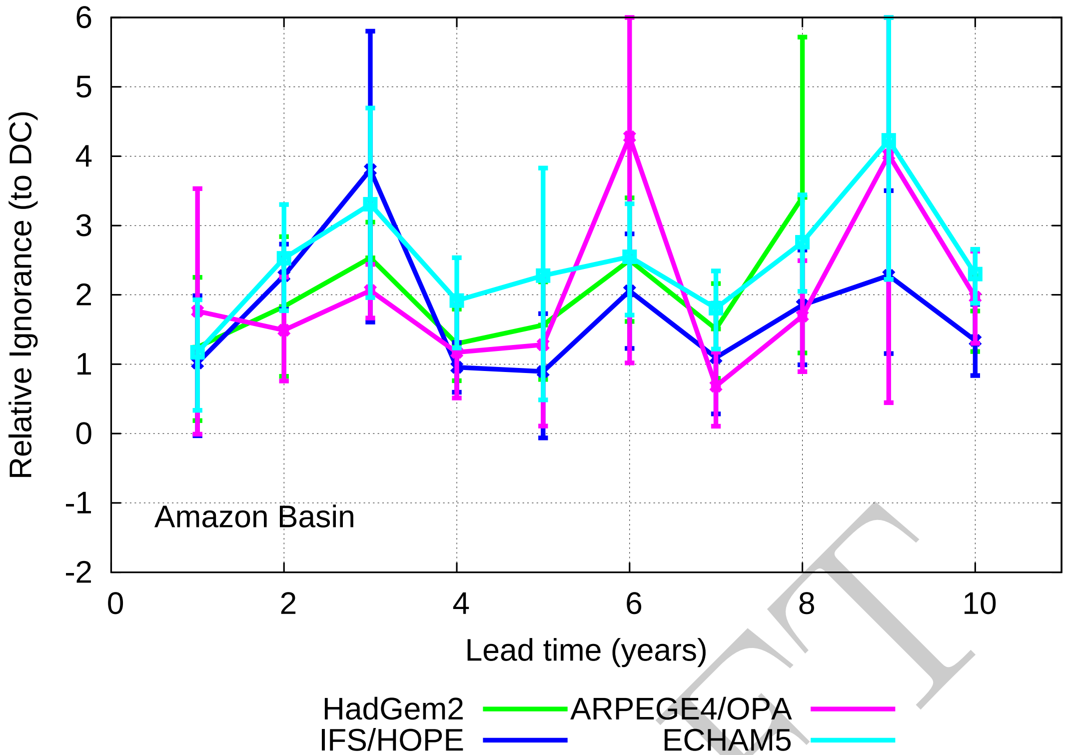

Fig. 6. Ignorance of the ENSEMBLES simulation models relative to the DC model for Amazon Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

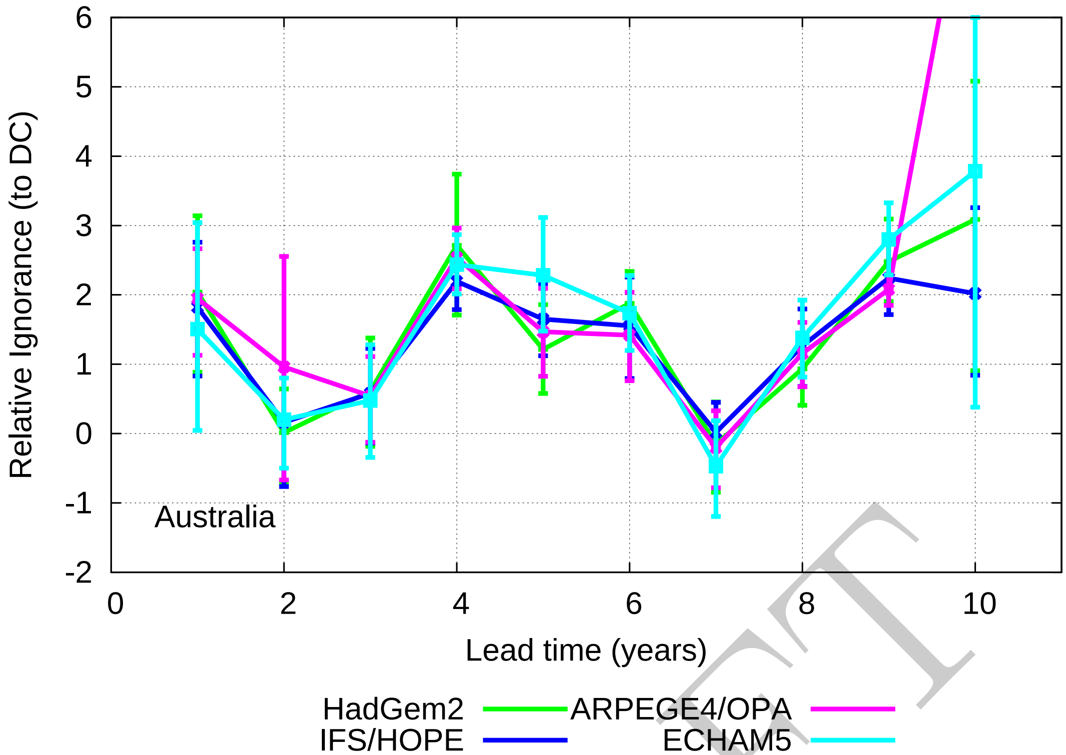

FIG. 7. Ignorance of the ENSEMBLES simulation models relative to the DC model for Australia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

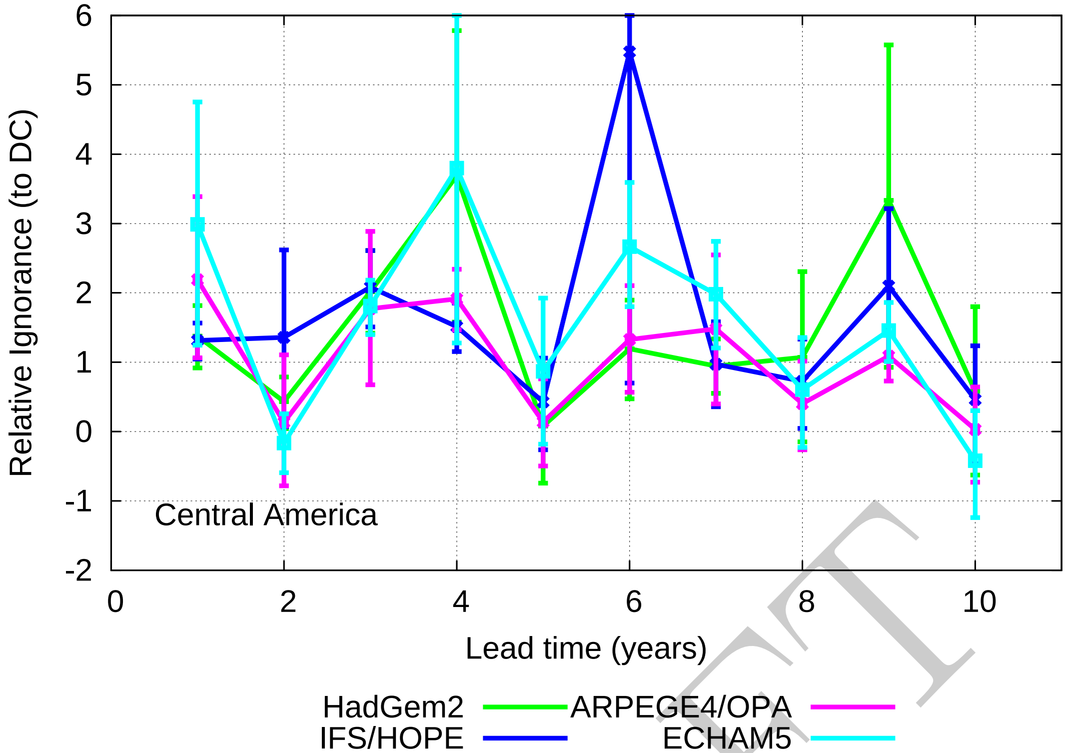

FIG. 8. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

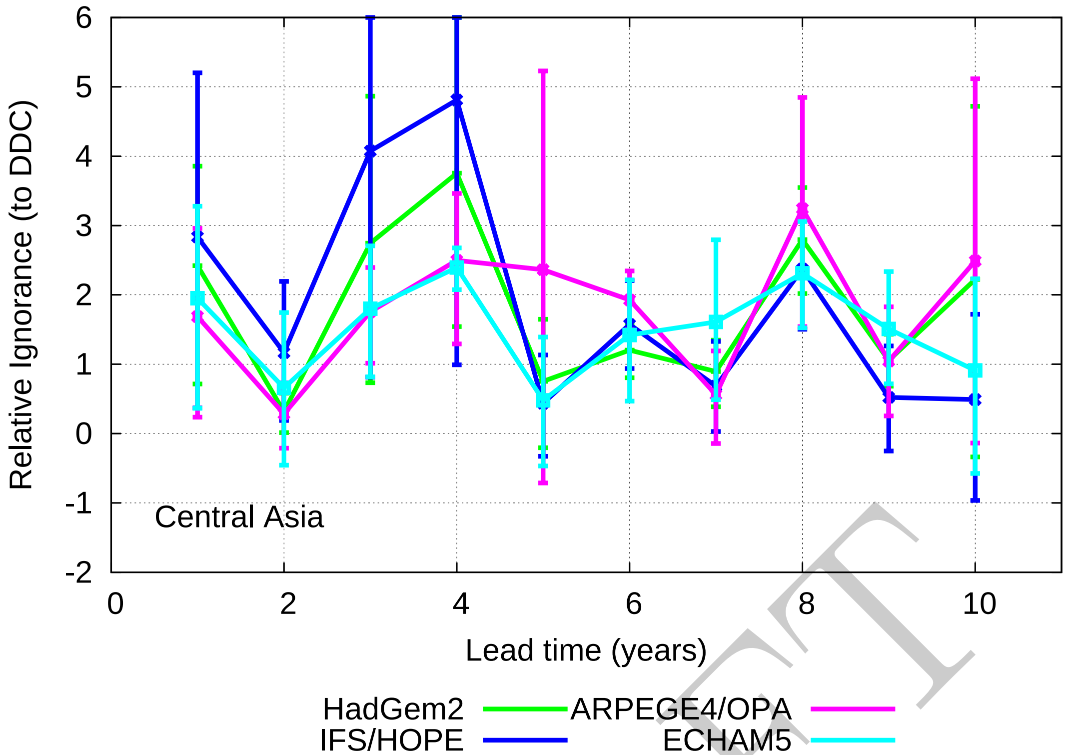

Fig. 9. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

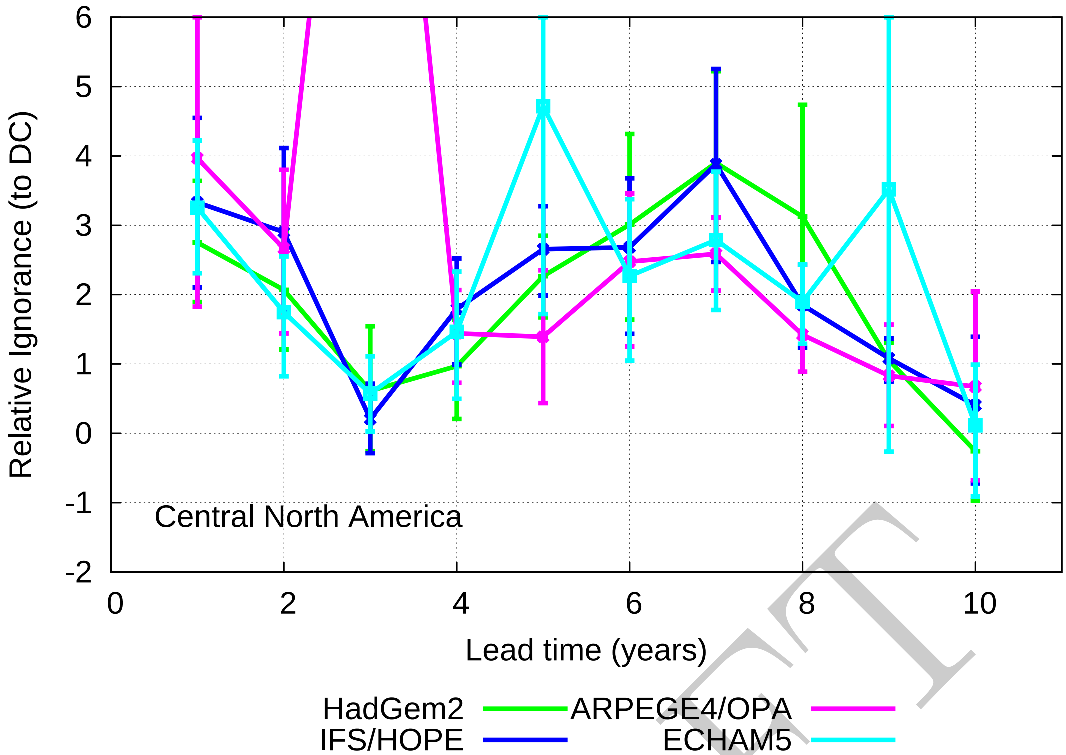

FIG. 10. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

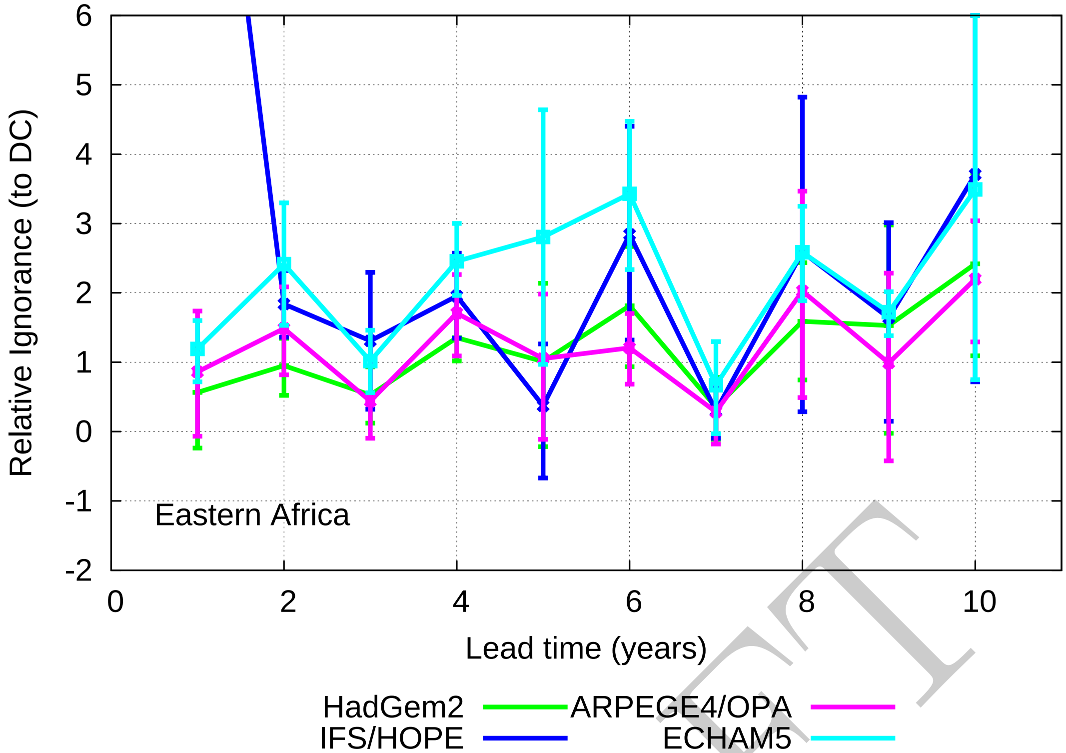

Fig. 11. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

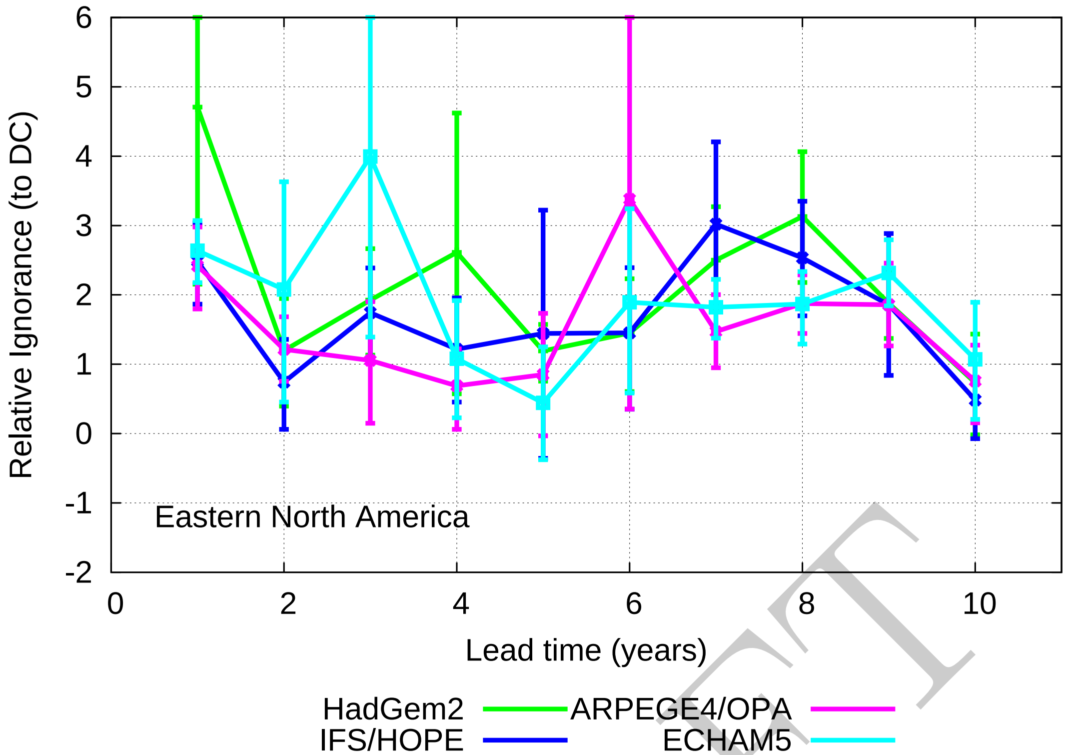

FIG. 12. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

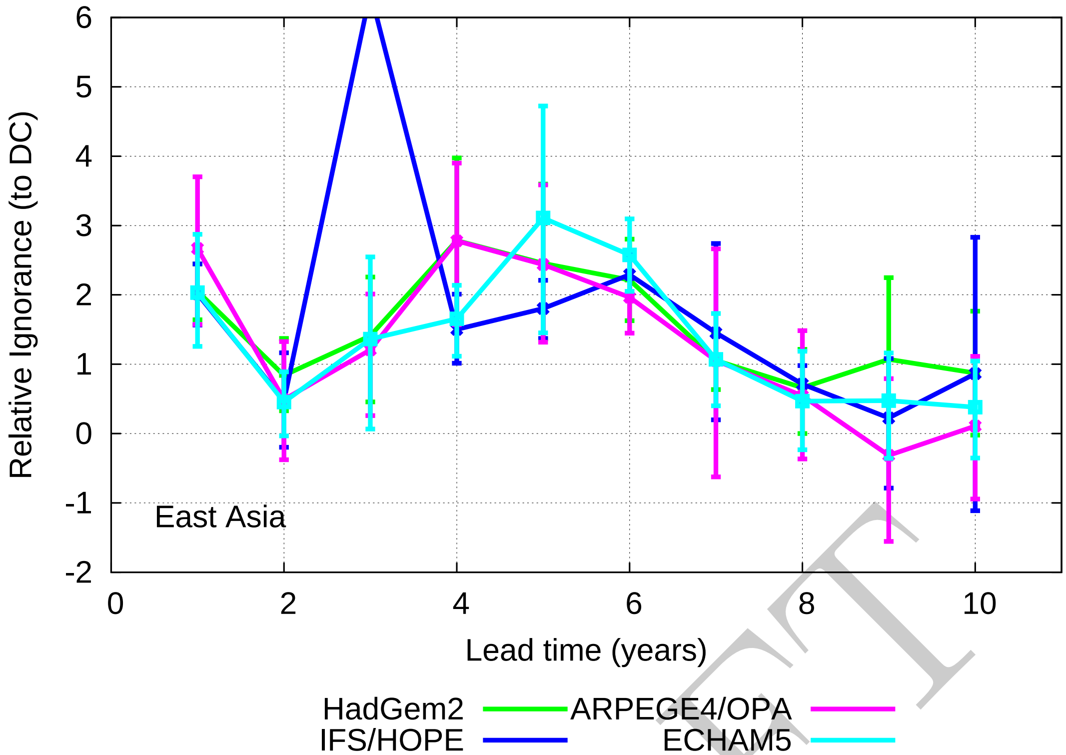

FIG. 13. Ignorance of the ENSEMBLES simulation models relative to the DC model for East Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

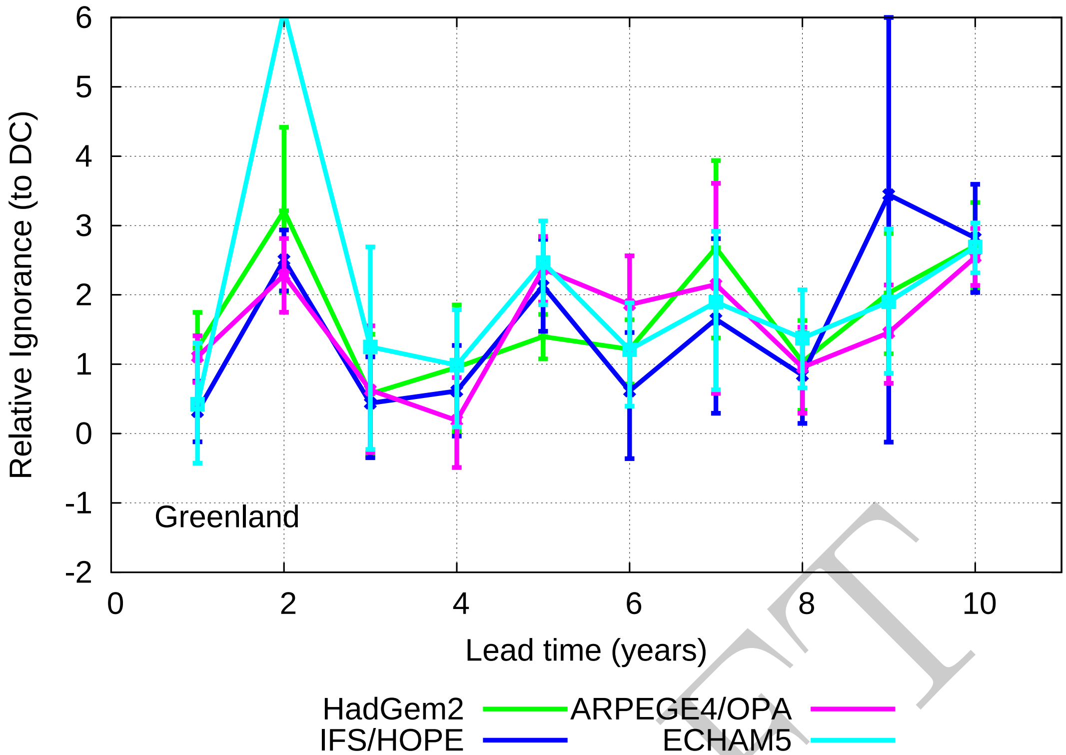

Fig. 14. Ignorance of the ENSEMBLES simulation models relative to the DC model for Greenland. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

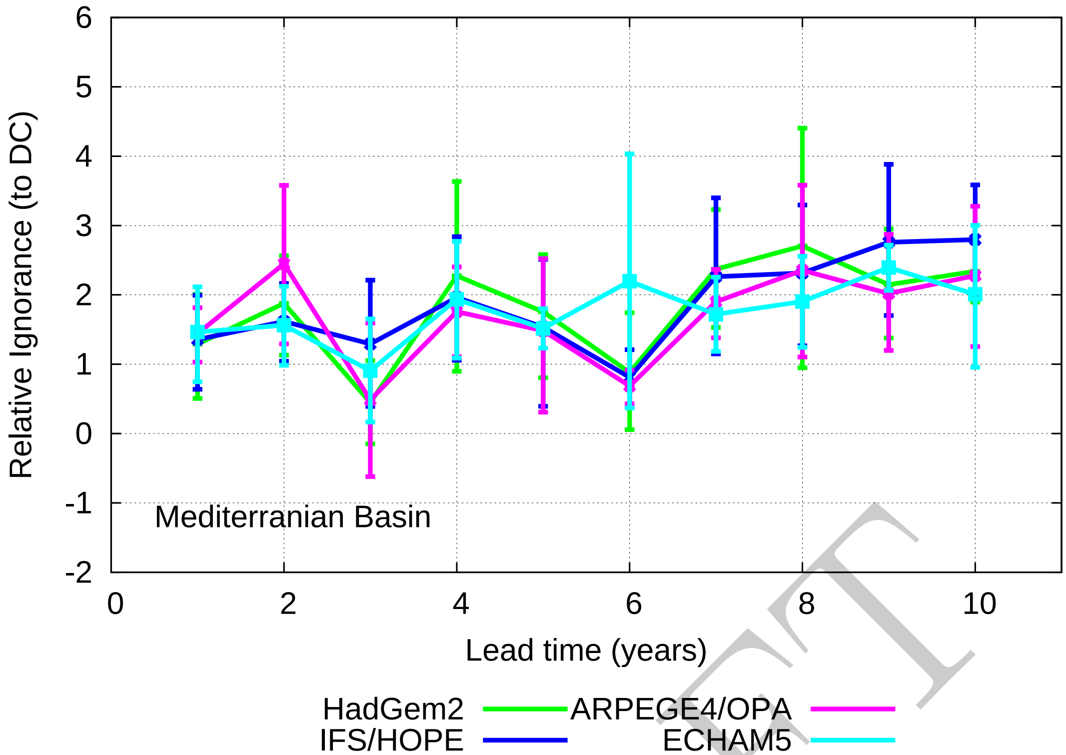

FIG. 15. Ignorance of the ENSEMBLES simulation models relative to the DC model for Mediterranian Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

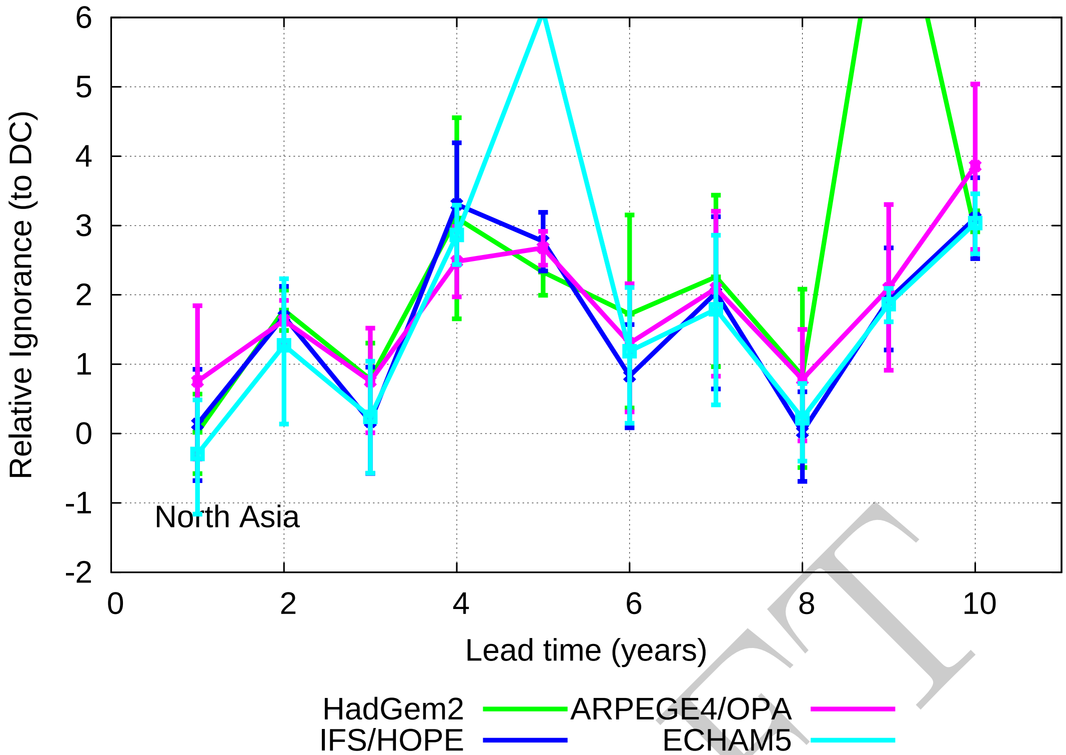

Fig. 16. Ignorance of the ENSEMBLES simulation models relative to the DC model for North Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

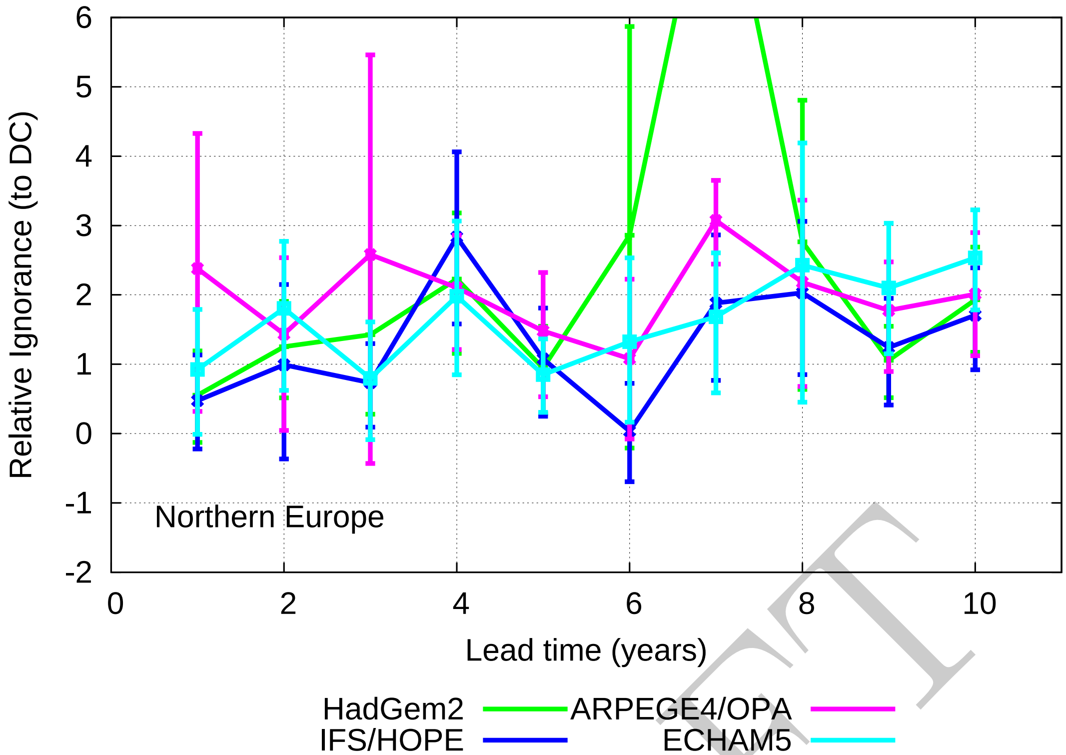

FIG. 17. Ignorance of the ENSEMBLES simulation models relative to the DC model for Northern Europe. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

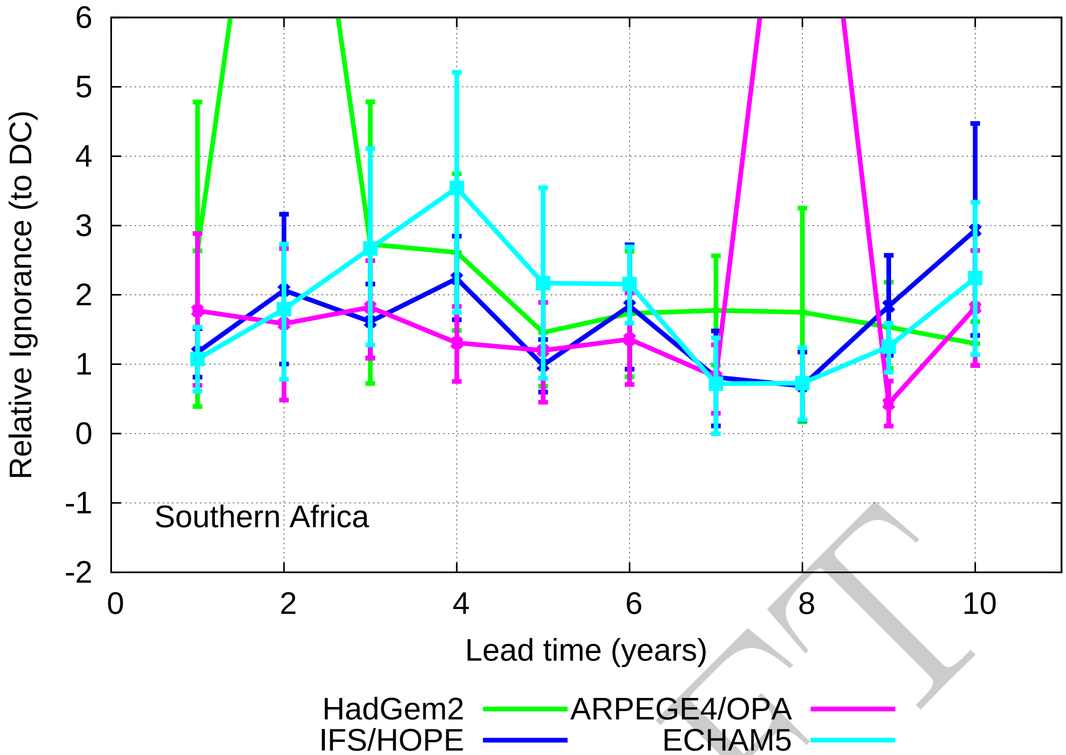

FIG. 18. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

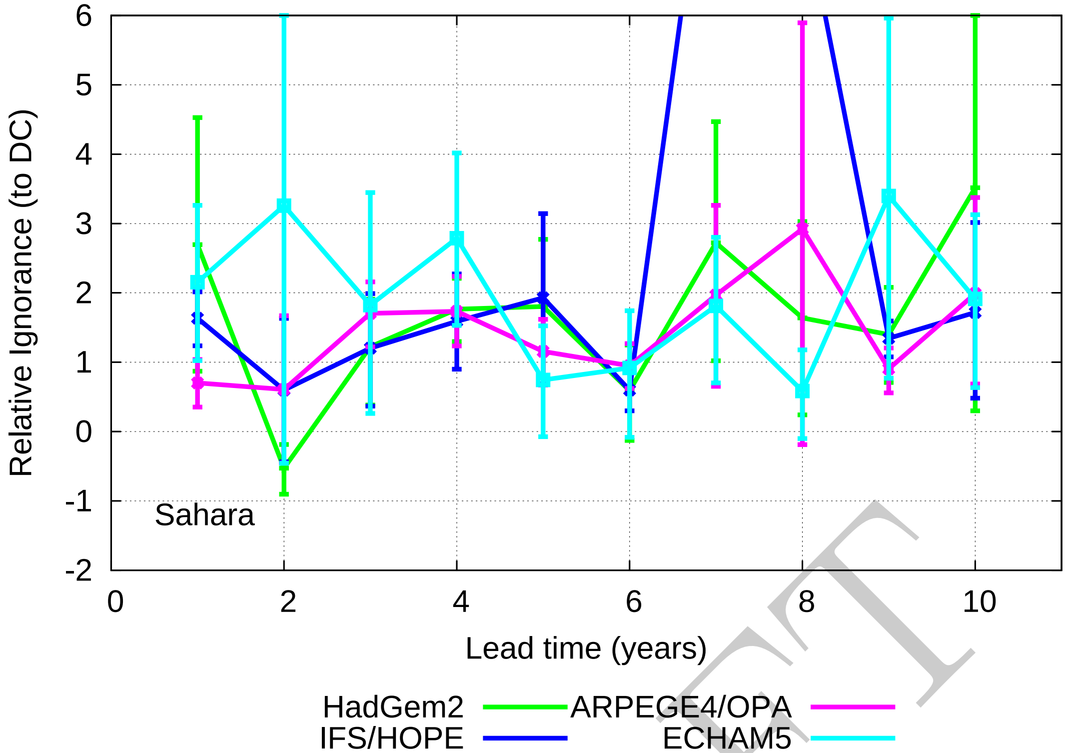

FIG. 19. Ignorance of the ENSEMBLES simulation models relative to the DC model for Sahara. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

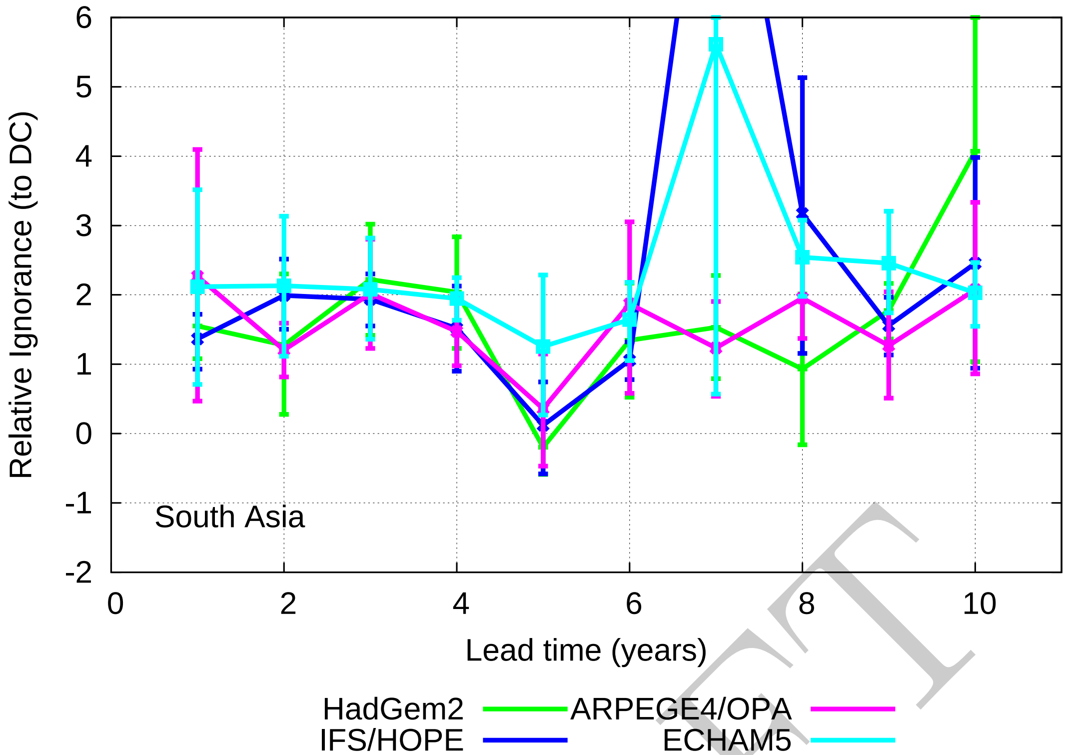

FIG. 20. Ignorance of the ENSEMBLES simulation models relative to the DC model for South Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

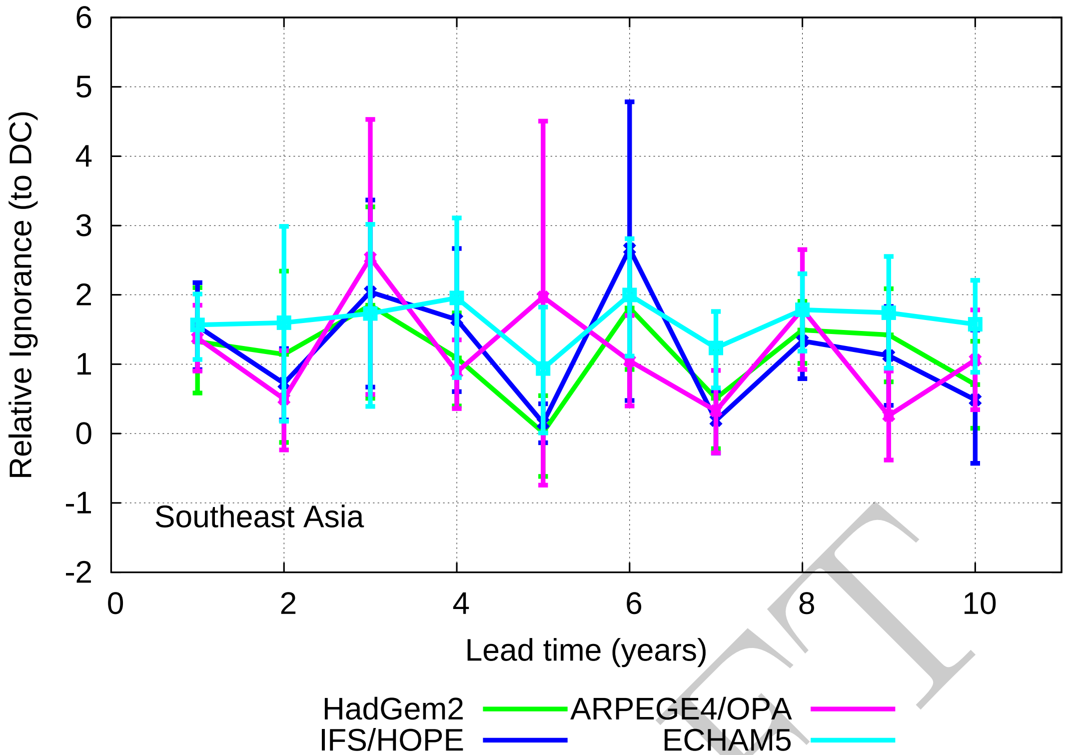

FIG. 21. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southeast Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

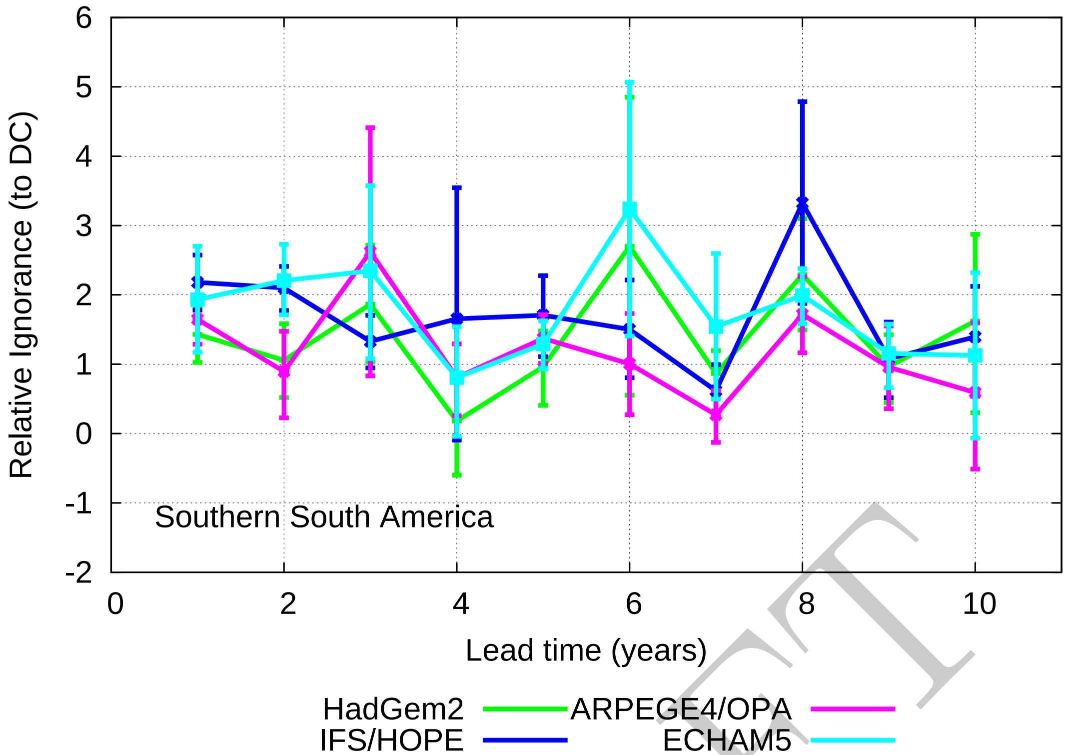

FIG. 22. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern South America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

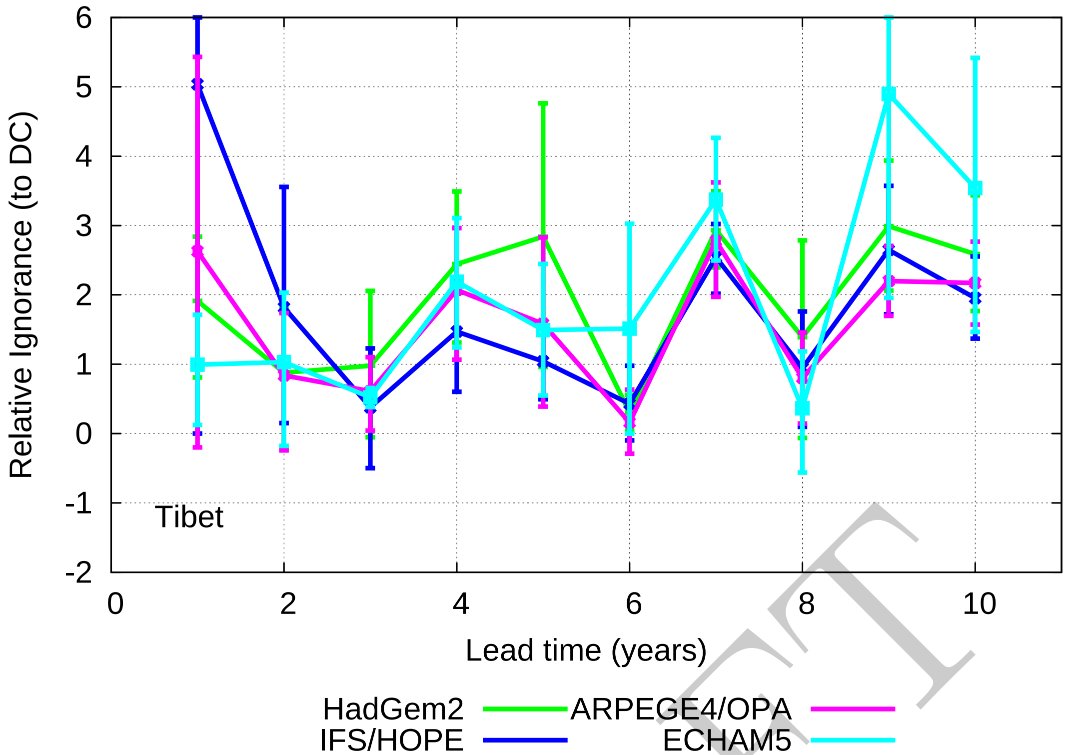

FIG. 23. Ignorance of the ENSEMBLES simulation models relative to the DC model for Tibet. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

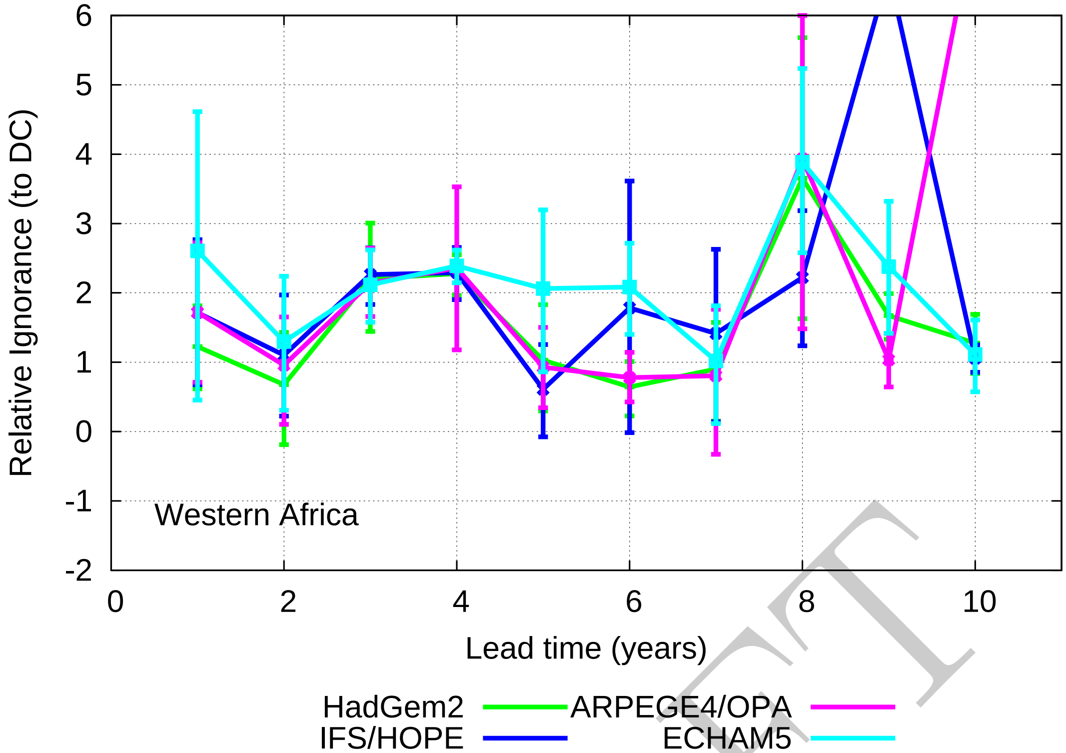

FIG. 24. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

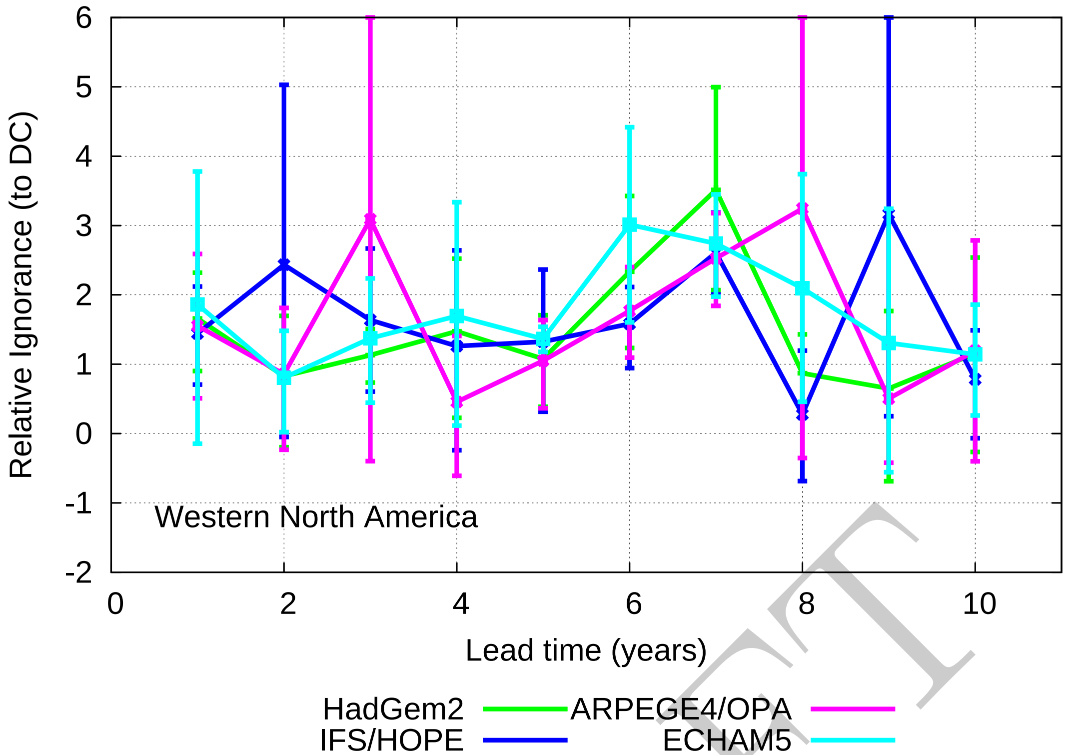

FIG. 25. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

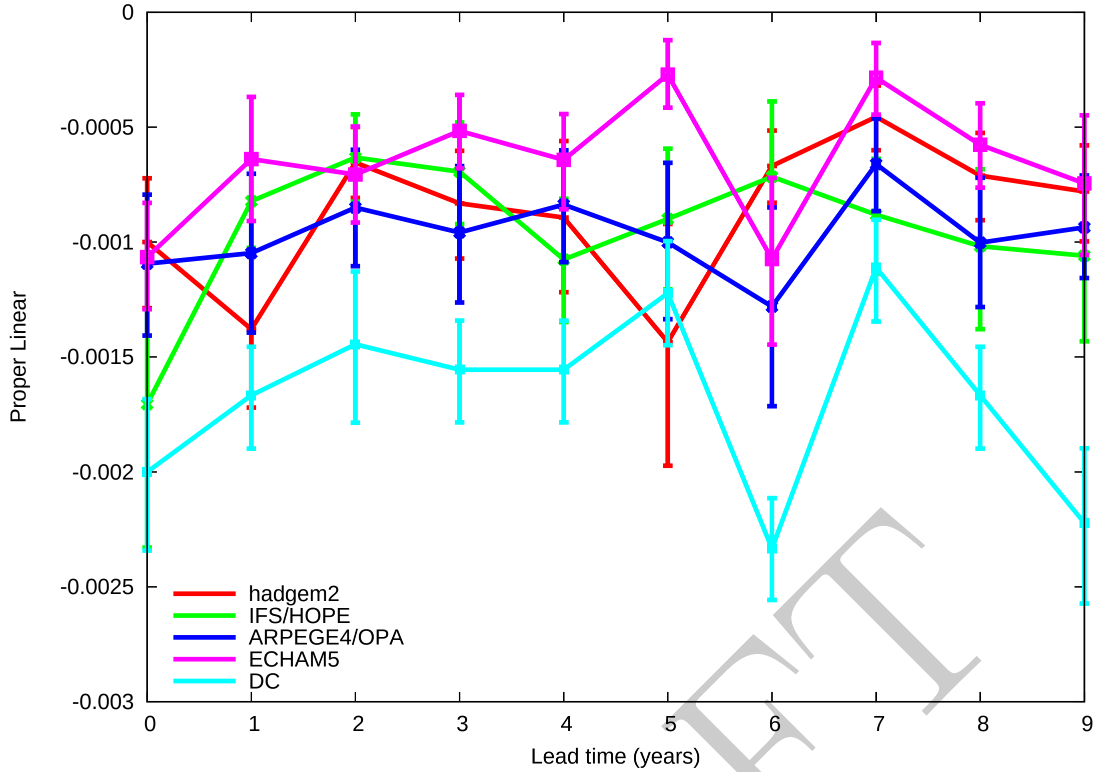

FIG. 26. Proper linear score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better foecasts. The DC model is shown to outperform the simulations models at most lead times.

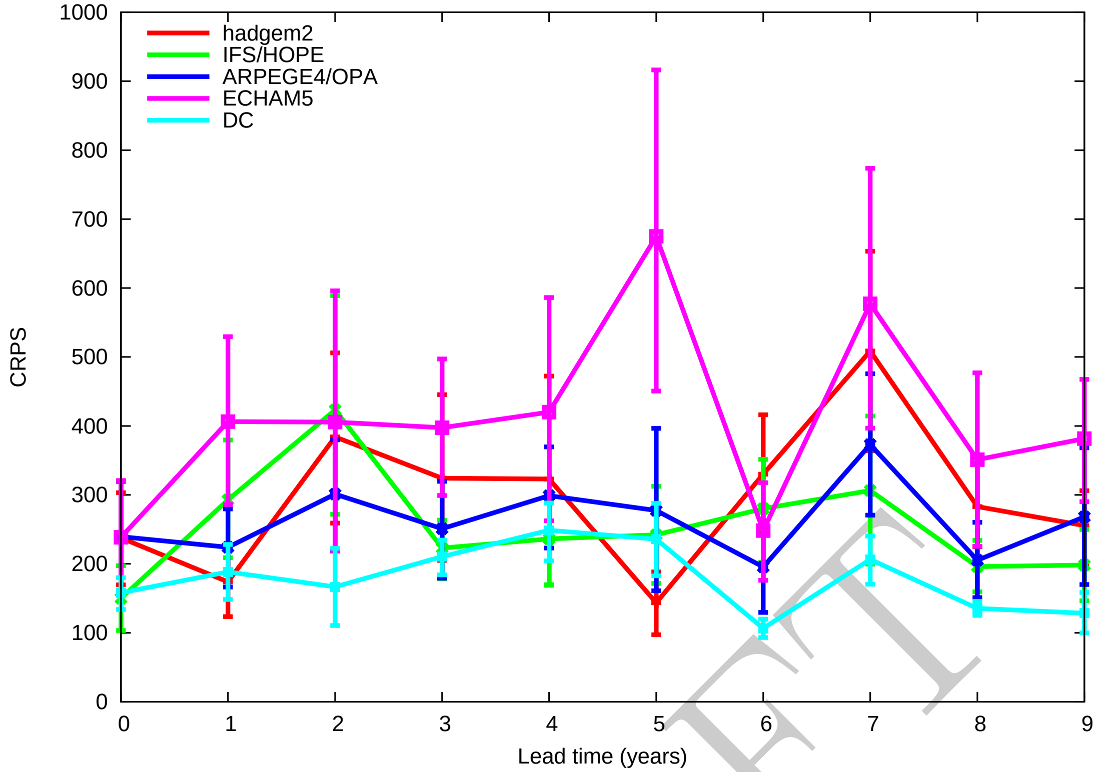

FIG. 27. CRPS score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better forecasts. The DC model is shown to outperform the simulations models at most lead times.

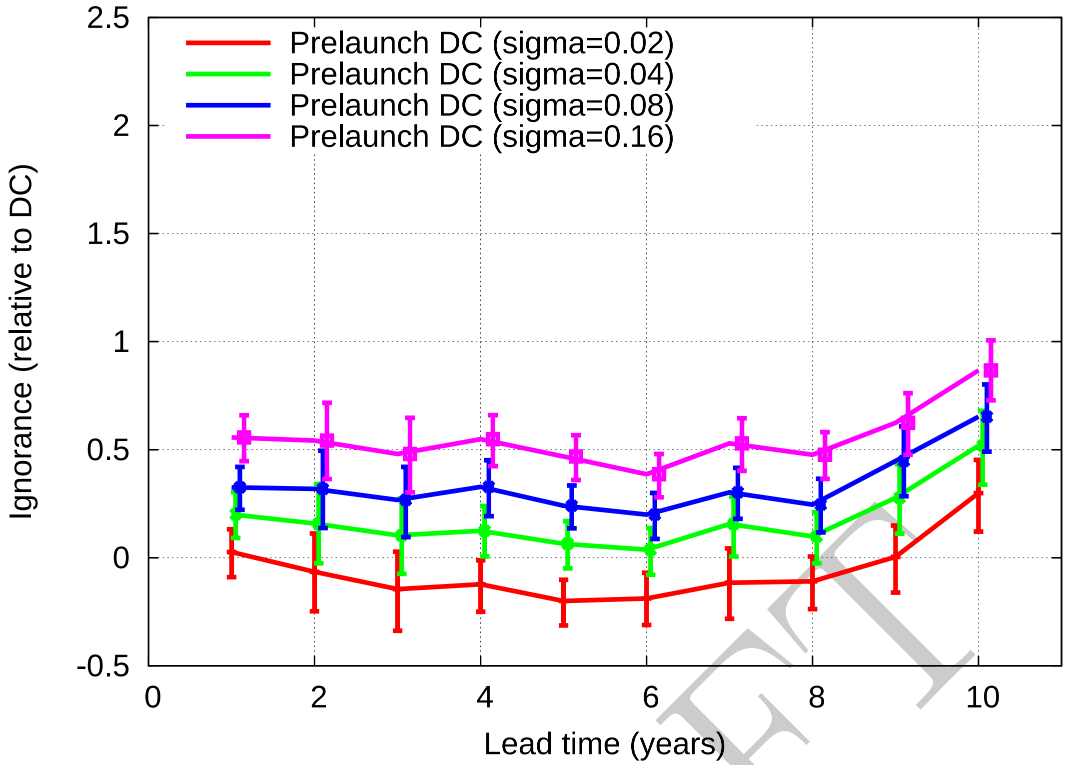

FIG. 28. Ignorance of the Prelaunch DC empirical model with kernel widths as labelled relative to the cross-validation DC model. Increasing the kernel width parameter from 0.02 to 0.16 results in a loss of skill of approximately half a bit, although for the kernel width value used in this paper (0.08) there is shown to be no significant loss of skill relative to the standard DC model.

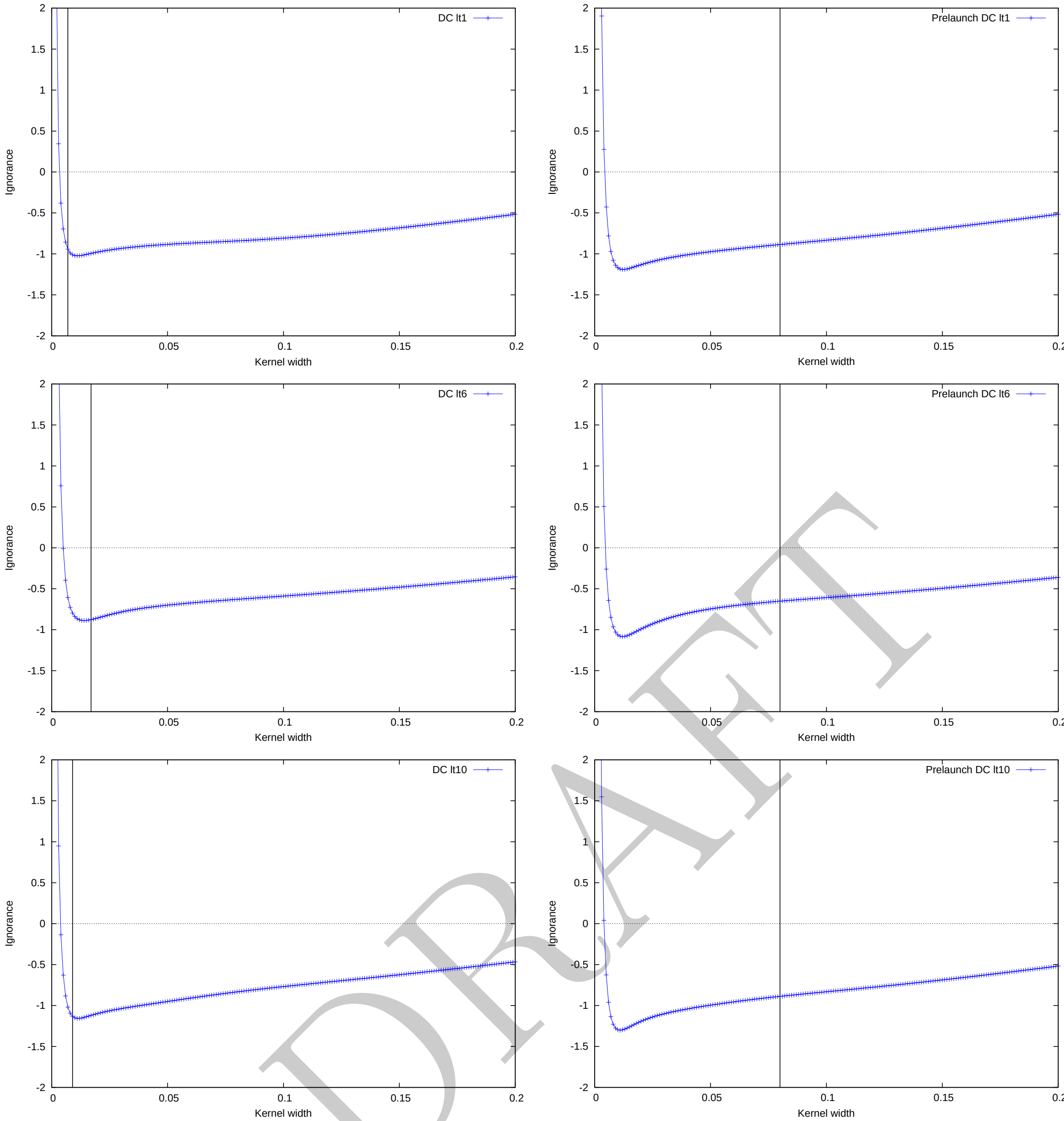

FIG. 29. Ignorance as a function of the kernel width parameter over the full set of hindcast simulations (*i.e.* with no cross-validation) for the DC (left panels) and Prelaunch DC (right panels) models at lead time one (a and b), six (c and d) and ten (e and f). The vertical bars in each case illustrate the kernel width parameters employed in the main manuscript. In the DC model parameters were attained through true-leave-one-out cross-validation. In the Prelaunch DC model a kernel spread value of 0.08 was chosen for comparison with DC and to test the robustness of the results to choices in the parameters for ensemble interpretation (although this value does not correspond to the lowest value of in-sample skill).

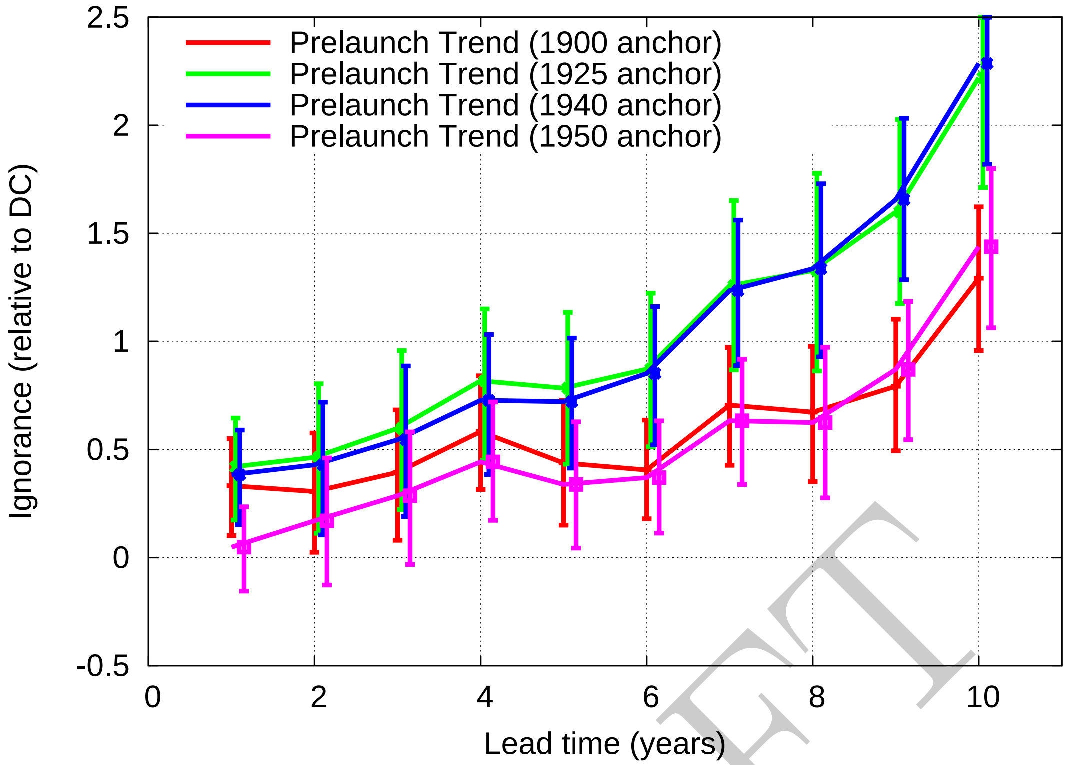

FIG. 30. Ignorance of the Prelaunch trend empirical model for different anchor times relative to the cross-validation DC model. Scores above zero indicate that DC outperforms the Prelaunch Trend model by up to half a bit at early lead times, and up to two bits (DC placing up to 4 times more probability on the observed outcome than the Prelaunch Trend model) up to ten years ahead, depending on the anchor year for the trend model.