Centre for Climate Change Economics and Policy

The Munich Re Programme: *Evaluating the Economics of Climate Risks and Opportunities in the Insurance Sector*

Grantham Research Institute on Climate Change and the Environment

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

Emma B. Suckling and Leonard A. Smith

21st October 2013

Centre for Climate Change Economics and Policy
Working Paper No. 167

Munich Re Programme Technical Paper No. 19

Grantham Research Institute on Climate Change and the Environment
Working Paper No. 150

The Grantham Foundation for the Protection of the Environment
Global Green Growth Institute
ESRC Economic & Social Research Council
Munich RE
University of Leeds
LSE The London School of Economics and Political Science

**The Centre for Climate Change Economics and Policy (CCCEP)** was established by the University of Leeds and the London School of Economics and Political Science in 2008 to advance public and private action on climate change through innovative, rigorous research. The Centre is funded by the UK Economic and Social Research Council and has five inter-linked research programmes:

1. Developing climate science and economics
2. Climate change governance for a new global deal
3. Adaptation to climate change and human development
4. Governments, markets and climate change mitigation
5. The Munich Re Programme - Evaluating the economics of climate risks and opportunities in the insurance sector (funded by Munich Re)

More information about the Centre for Climate Change Economics and Policy can be found at: http://www.cccep.ac.uk.

**The Munich Re Programme** is evaluating the economics of climate risks and opportunities in the insurance sector. It is a comprehensive research programme that focuses on the assessment of the risks from climate change and on the appropriate responses, to inform decision-making in the private and public sectors. The programme is exploring, from a risk management perspective, the implications of climate change across the world, in terms of both physical impacts and regulatory responses. The programme draws on both science and economics, particularly in interpreting and applying climate and impact information in decision-making for both the short and long term. The programme is also identifying and developing approaches that enable the financial services industries to support effectively climate change adaptation and mitigation, through for example, providing catastrophe insurance against extreme weather events and innovative financial products for carbon markets. This programme is funded by Munich Re and benefits from research collaborations across the industry and public sectors.

**The Grantham Research Institute on Climate Change and the Environment** was established by the London School of Economics and Political Science in 2008 to bring together international expertise on economics, finance, geography, the environment, international development and political economy to create a world-leading centre for policy-relevant research and training in climate change and the environment. The Institute is funded by the Grantham Foundation for the Protection of the Environment and the Global Green Growth Institute, and has five research programmes:

1. Global response strategies
2. Green growth
3. Practical aspects of climate policy
4. Adaptation and development
5. Resource security

More information about the Grantham Research Institute on Climate Change and the Environment can be found at: http://www.lse.ac.uk/grantham.

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

Emma B. Suckling and Leonard A. Smith

Centre for the Analysis of Time Series,

London School of Economics, Houghton Street,

London, WC2A 2AE, UK

Tel: +44 207 955 6015, Email: cats@lse.ac.uk

21st October 2013



ABSTRACT

While state-of-the-art models of the Earth's climate system have improved tremendously over the last twenty years, nontrivial structural flaws still hinder their ability to forecast the decadal dynamics of the Earth system realistically. Contrasting the skill of these models not only with each other but also with empirical models can reveal the space and time scales on which simulation models exploit their physical basis effectively and quantify their ability to add information to operational forecasts. The skill of decadal probabilistic hindcasts for annual global-mean and regional-mean temperatures from the EU ENSEMBLES project is contrasted with several empirical models. Both the ENSEMBLES models and a "Dynamic Climatology" empirical model show probabilistic skill above that of a static climatology for global-mean temperature. The Dynamic Climatology model, however, often outperforms the

ENSEMBLES models. The fact that empirical models display skill similar to that of today's state-of-the-art simulation models suggests that empirical forecasts can improve decadal forecasts for climate services, just as in weather, medium range, and seasonal forecasting. It is suggested that the direct comparison of simulation models with empirical models becomes a regular component of large model forecast evaluations. Doing so would clarify the extent to which state-of-the-art simulation models provide information beyond that available from simpler empirical models and clarify current limitations in using simulation forecasting for decision-support. Ultimately the skill of simulation models based on physical principles is expected to surpass that of empirical models in a changing climate; their direct comparison provides information on progress toward that goal which is not available in model-model intercomparisons.



# 1. Introduction

State-of-the-art dynamical simulation models of the Earth's climate system[1] are often used to make probabilistic predictions about the future climate and related phenomena with the aim of providing useful information for decision support (Anderson et al. 1999; UK-MetOffice 2011; Weigela and Bowlerb 2009; Alessandri et al. 2011; Hagedorn et al. 2005; Hagedorn and Smith 2009; Meehl et al. 2009; Doblas-Reyes et al. 2010, 2011; IPCC 2007; Reifen and Toumi 2009). Evaluating the performance of such predictions from a model, or set of models, is crucial not only in terms of making scientific progress, but also in determining how much information may be available to decision-makers via climate services.

[1]Models that use physical principles to simulate the Earth's climate are often called general circulation models (GCMs), coupled atmosphere-ocean global climate models (AOGCMs) or Earth system models (ESMs). Such models are referred to as simulation models throughout this paper. The key distinction is their explicit use of physical principles to simulate the system of interest. Simulation models are to be contrasted with models based almost solely on observations, which are hereafter referred to as 'empirical models' following (van den Dool 2007)

It is desirable to establish a robust and transparent approach to forecast evaluation, for the purpose of examining the extent to which today's best available models are adequate over the spatial and temporal scales of interest for the task at hand. A useful reality check is provided by comparing the simulation models not only with other simulation models, but also with empirical models which do not include direct physical simulation.

Decadal prediction brings several challenges for the design of ensemble experiments and their evaluation (Meehl et al. 2009; van Oldenborgh et al. 2012; Doblas-Reyes et al. 2010; Fildes and Kourentzes 2011; Doblas-Reyes et al. 2011); the analysis of decadal prediction systems will form a significant focus of the IPCC's fifth assessment report (AR5). Decadal forecasts are of particular interest both for information on the impacts over the next ten years, as well as from the perspective of climate model evaluation. Hindcast experiments over an archive of historical observations allow approaches from empirical forecasting to be used for model evaluation. Such approaches can aid in the evaluation of forecasts from simulation models (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012) and potentially increase the practical value of such forecasts through blending forecasts from simulation models with forecasts from empirical models that do not include direct physical simulation (Bröcker and Smith 2008).

This paper contrasts the performance of decadal probability forecasts from simulation models with that of empirical models constructed from the record of available observations. Empirical models are unlikely to yield realistic forecasts for the future once climate change moves the Earth system away from the conditions observed in the past. A simulation model, which aims to capture the relevant physical processes and feedbacks, is expected to be at least competitive with the empirical model. If this is not the case in the recent past, then it is reasonable to demand evidence that those particular simulation models are likely to be more informative than empirical models in forecasting the near future.

A set of decadal simulations from the ENSEMBLES experiment (Hewitt and Griggs 2004; Doblas-Reyes et al. 2010), a precursor to the Coupled Model Intercomparison Project Phase

5 (CMIP5) decadal simulations (Taylor et al. 2009) is considered. The ENSEMBLES probability hindcasts are contrasted with forecasts from empirical models of the static climatology, persistence and a "Dynamic Climatology" model developed for evaluating other dynamical systems (Smith 1997; Binter 2011). Ensemble members are transformed into probabilistic forecasts via kernel dressing (Bröcker and Smith 2008) and their quality quantified according to several proper scoring rules (Bröcker and Smith 2006). The ENSEMBLES models do not demonstrate significantly greater skill than that of an empirical Dynamic Climatology model either for global mean temperature or for the land-based Giorgi region[2] temperatures (Giorgi 2002).

It is suggested that the direct comparison of simulation models with empirical models become a regular component of large model forecast evaluations. The methodology is easily adapted to other climate forecasting experiments and can provide a useful guide to decision-makers about whether state-of-the-art forecasts from simulation models provide additional information to that available from easily constructed empirical models.



An overview of the ENSEMBLES models used for decadal probabilistic forecasting is discussed in section 2. The appropriate choice of empirical model for probabilistic decadal predictions forms the basis of section 3, while section 4 contains details of the evaluation framework and the transformation of ensembles into probabilistic forecast distributions. The performance of the ENSEMBLES decadal hindcast simulations is presented in section 5 and compared to that of the empirical models. Section 6 then provides a summary of conclusions and a discussion of their implications. The Supplementary Material includes graphics for models not shown in the main text, comparisons with alternative empirical models, results for regional forecasts and the application of alternative (proper) skill scores. The basic conclusion is relatively robust: the empirical Dynamic Climatology (DC) model

[2]Giorgi regions are a set of land-based regions, defined in terms of simple rectangular areas and chosen based on a qualitative understanding of current climate zones and on judgements about the performance of climate models within these zones.

often outperforms the simulation models in terms of probability forecasting of temperature.

# 2. Decadal prediction systems

Given the timescales required to obtain fresh out-of-sample observations for the evaluation of decadal forecast systems, forecast evaluation is typically performed in-sample using hindcasts. Hindcasts (or retrospective forecasts) are predictions made as if they had been launched on dates in the past, and allow some comparison of model simulations with observations. Of course, simulation models have been designed after the study of this same historical data, so their ability to reproduce historical observations carries significantly less weight than success out-of-sample. Failure in-sample, however, can be instructive.

In a changing climate, even out-of-sample skill is no guarantee of future performance, due to the nonlinear nature of the response to external forcing (Smith 2002; Reifen and Toumi 2009; IPCC 2007). Nevertheless, the fact that only simulation models based on the appropriate physical principles are expected to be able to generalize to new physical conditions provides no evidence that today's state-of-the-art simulation models can do so. Contrasting probability forecasts from simulation models with those from empirical models is one guide to gauging the additional information derived from the physical basis of the simulation model-based forecasts. In practice, the most skillful probability forecast is often based on combining the information from both simulation models and empirical models (van den Dool 2007; Hoeting et al. 1999; Unger et al. 2009; Bröcker and Smith 2008; UK-MetOffice 2011).

Decadal predictions aim to accurately represent both the intrinsic variability and forced response to changes in the Earth system (Meehl et al. 2009). Decadal simulation models now assimilate observations of the current state of the Earth system as initial conditions in the model (Pierce et al. 2004; Troccoli and Palmer 2007)[3]. At present it is not clear

[3] In reality, of course, no such distinct entities exist given the nonlinearity of the Earth System. The

whether initialising the model with observations at each forecast launch improves the skill of decadal forecasts (Pohlmann et al. 2009; Hawkins et al. 2011; Smith et al. 2007; Keenlyide et al. 2008; Smith et al. 2010; van Oldenborgh et al. 2012; Kim et al. 2012). At a more basic level, the ability to provide useful decadal predictions using simulation models is yet to be firmly established. Probabilistic hindcasts, based on simulations from Stream 2 of the ENSEMBLES project (further details of which can be found in (Doblas-Reyes et al. 2010) and in the Appendix), do not demonstrate significantly more skill than that of simple empirical models.

Figure 1 illustrates the 2-year running mean of simulated global mean temperature from the four simulation models in the multi-model ensemble experiment of the ENSEMBLES project over the full set of decadal hindcasts. Observations from the HadCRUT3 dataset and ERA40 reanalysis are shown for comparison. HadCRUT3 is used as the outcome archive for both the model evaluation and construction of the empirical model. Using ERA40 for the verification instead of HadCRUT3 does not change the conclusions about the model skill significantly (results not presented here). Global mean temperature is chosen for the analysis as simulation models are expected to perform better over larger spatial scales (IPCC 2007). Even at the global scale the raw simulation output is seen to differ from the observations both in terms of absolute values, as well as in dynamics. Three of the four models display a substantial model drift away from the observed global mean temperature, the ECHAM5 model is the exception. The fact that some of the models exhibit a substantial drift, but not others, reflects the fact that different models employ different initialisation schemes (Keenlyide et al. 2005). ECHAM5 both assimilates anomalies and forecasts anomalies. Assimilating anomalies is intended to reduce model drift[4] (Pierce et al. 2004); the remaining models are initialised from observed conditions.



nature of "intrinsic variability" is inextricably linked to the state of the Earth System; there is no separation into a natural component and a forced component.

[4]For point forecasts, forecasting anomalies allows an immediate apparent bias reduction at short lead times on the order of the model's systematic error.

A standard practice for dealing with model drift is to apply an empirical (linear) "bias correction" to the simulation runs (Stockdale 1997; Jolliffe and Stephenson 2003). Such a procedure both assumes that the bias of a given model at a given lead time does not change in the future, and is expected to break the connection between the underlying physical processes in the model and its forecast. Bias correction is often applied using the (sample) mean forecast error at each forecast lead time. The mean forecast error is shown as a function of lead time for global mean temperature in figure 2 for each of the ENSEMBLES models. Here, lead time 1 indicates the average of the first 12 months of each simulation, initialised in November of the launch year.

The focus in this paper is on probability forecasts, specifically on contrasting the skill of simulation model probability forecasts with empirical model probability forecasts. On weather forecast timescales and in the medium range, simulation model based probability forecasts clearly have more skill than empirical model probability forecasts based on climatology (Hagedorn and Smith 2009). The question is whether, in the context of decadal probability forecasting, simulation models produce decadal probability predictions that are more skillful than simple empirical models. Answering this question requires defining an appropriate empirical model.

# 3. Empirical models for decadal prediction

Empirical models are common in forecast evaluation (Barnston et al. 1994; Colman and Davey 2003; van Oldenborgh et al. 2005, 2012; Lee et al. 2006; van den Dool 2007; Laepple et al. 2008; Krueger and von Storch 2011; Wilks 2011). They are used to quantify the information a simulation model adds beyond the naïve baseline the empirical models define. They have also been used to estimate forecast uncertainty (Smith 1992), both as benchmarks for simulation forecasts, and as a source of information to be combined with simulation model forecasts (van den Dool 2007; Unger et al. 2009; Smith 1997; UK-MetOffice 2011; Hagedorn and Smith 2009).

Empirical models based on historical observations cannot be expected to capture previously unobserved dynamics. Two empirical models typically used in forecast evaluation are the climatological distribution and the persistence model. In the analysis below, a static climatology defines a probabilistic distribution generated through the kernel dressing and cross-validation procedures applied to the observational record (Bröcker and Smith 2008; Hoeting et al. 1999), as outlined in section 4. Persistence forecasts are defined according to a similar procedure, based on the last observation, persisted as a single ensemble member for each launch. These models are not expected to prove ideal in a changing climate, nevertheless information regarding the ability (or inability) of a simulation model to outperform these simple empirical models is of value. Alternative empirical models for probability forecasts, more appropriate for a changing climate, define a Dynamic Climatology based on ensemble random analogue prediction (eRAP) (Smith 1997; Paparella et al. 1997). Empirical forecasts are also used as benchmarks for evaluating point forecasts of decadal climate predictions (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012; Doblas-Reyes et al. 2010).

Analogue forecasting uses the current state (perhaps with other recent states (Smith 1997)) to define analogues within the observational record (van den Dool 1994; Lorenz 1963; van den Dool 2007). A distribution based on images of each analogue state (the observation immediately following the analogue state) then defines the ensemble forecast. Analogues may be defined in a variety of ways, including near neighbours either in observation space or in a delay reconstruction (Smith 1994, 1997). The ensemble members may be formed using the complete set of available analogue states (Dynamic Climatology (Smith 1997; Binter 2011)), or by selecting from the nearest neighbours at random, with the probability of selecting a particular neighbour related to the distance (in the state space) between the prediction point and the neighbour (the random analogue prediction method (Paparella et al. 1997)).

The dynamic climatologies constructed below provide $l$-step ahead forecast distributions based on the current state and differences defined in the observational record. There are two approaches to forming such a Dynamic Climatology: (i) direct and (ii) iterated (Smith 1992). The direct Dynamic Climatology (DC) approach used below considers the $l$-step differences in the observational record (for example a 1-step difference might be the temperature difference between the current state and its immediately preceding state). A distribution is formed for each value of $l$ from the corresponding differences using all the observations after some start date, thus the size of the ensemble decreases linearly with lead time due to the finite size of the archive. For a forecast of a scalar quantity, such as the global mean temperature below, the DC ensemble at lead time $l$ launched at time $t$ consists of the set of $N_l$ values:

$$e_i = S_t +^l \Delta_i,\ i = 1, ..., N_l, \tag{1}$$

where $S_t$ is the initial condition at time $t$ and ${}^l\Delta_i, i = 1, ..., N_l$ is the set of $l^{th}$ differences in the observational record. Figure 3 illustrates the DC model for global mean temperature, launched at five-year intervals, as in the ENSEMBLES hindcasts. A true out-of-sample forecast up to the year 2015, initialised to the observed global mean temperature in 2004, is also included. Each lead time 1 forecast is based on an ensemble of 48 members. In real-time forecasting $N_l = N - l$, while for cross-validation purposes the ensembles in figure 3 use $N - l - 1$, omitting the $\Delta_j$ corresponding to the year being forecast. Thus at lead time 9 each forecast is based on an ensemble of 40 members. The DC approach is shown below to outperfom the ENSEMBLES models when forecasting global mean temperature.

# 4. Probability forecasts from ensembles

No forecast is complete without an estimate of forecast skill (Tennekes et al. 1987). Probability forecasts allow a complete description of the skill from an ensemble prediction system; they may be formed in several ways. The IPCC AR4 (IPCC 2007), for example, defines a likely range subjectively, applying the 'sixty-forty' rule[5] to the mean of the CMIP3 model global mean temperatures in 2100. Insofar as the forecast-outcome archive is larger for decadal timescales, objective statistical approaches are more easily deployed.

Decadal probability forecasts are formed by transforming the ensemble into a probability distribution function via kernel dressing (Bröcker and Smith 2008). A number of methods for this transformation exist and a selection will impact the skill of the forecast. The kernel dressed forecast based on an ensemble with $N$ members is (Bröcker and Smith 2008):

$$p(y : x, \sigma) = \frac{1}{N\sigma} \sum_{i=1}^{N} K\left(\frac{y - (x^i + \mu)}{\sigma}\right), \tag{2}$$

where $x^i$ is the $i^{th}$ ensemble member, $\mu$ is the offset of the kernel mean (this offset may have a different value than the traditional "bias" term[6]) and $\sigma$ is the kernel width. In this paper, the kernel, $K$, is taken to be a Gaussian function,

$$K(\varepsilon) = \frac{1}{\sqrt{(2\pi)}} exp\left(-\frac{1}{2}\varepsilon^2\right). \tag{3}$$

The kernel parameters are fitted by minimising a chosen skill score (Jolliffe and Stephenson 2003) while avoiding information contamination[7].

The forecasts below are evaluated using the Ignorance score (Good 1952), defined as

$$S(p(y), Y) = -log_2(p(Y)), \tag{4}$$

[5] In chapter 10.5.4.6 of the AR4 (IPCC 2007) the "likely" range of global temperatures in 2100 are provided for each of several scenarios. Each range falls "within 40 to +60% of the multi-model AOGCM mean warming simulated for each scenario." (pg 810). Similar results are shown in figure 5 of the Summary for Policy Makers.

[6] Kernel dressing and blending aim to provide good probability forecasts; this goal need not coincide with minimizing the point forecast error of the ensemble mean.

[7] Information contamination occurs when critical information is used in a hindcast which would not have been available for a forecast actually made on the same launch date. While such contamination can never be eliminated completely if the historical data is known, principled use of cross-validation can reduce its likely impact.

where $p(Y)$ is the probability assigned to the verification, $Y$. By convention the smaller the score the more skillful the forecast (Jolliffe and Stephenson 2003).

To contrast the skill of probability forecasts from two forecast systems it is useful to consider the relative Ignorance. The mean relative Ignorance of model 1 relative to model 2 is defined as

$$\begin{aligned} S_{rel}(p_1(y), p_2(y), Y) &= \frac{1}{F}\sum_{i=1}^{F} -log_2\left[\frac{p_1(Y_i)}{p_2(Y_i)}\right] \\ &= S(p_1(y), Y) - S(p_2(y), Y). \end{aligned} \tag{5}$$

If $p_2$ is taken as a reference forecast, then $S_{rel}$ defines 'zero skill' in the sense that $p_2$ will have $S_{rel} = 0$.

Appropriate reference forecasts will depend on the task at hand: they may include a static climatological distribution, a dynamic climatology, another simulation model or empirical model. The relative Ignorance quantifies (in terms of bits) the additional information provided by forecasts from one model, above that of the reference. A relative Ignorance score of $S_{rel}(p_1(y), p_2(y), Y) = -1$ means that the model forecast places, on average, twice (that is $2^1$) the probability mass on the verification than the reference forecast. Similarly, a score of
$S_{rel}(p_1(y), p_2(y), Y) = -1/2$ means $\sim$41% (that is $2^{1/2}$) more probability mass on average. In section 5, the static climatology, a persistence forecast and the DC model are chosen as references to measure performance against the ENSEMBLES simulation models. The parameters used to construct each empirical model forecast are each estimated under true-cross-validation: the forecast target decade is omitted from consideration.

The ENSEMBLES forecast-outcome archive contains at most nine forecast-outcome pairs. That is, there are only nine forecast launch dates, each with a maximum lead time of ten years. Outside of true out-of-sample evaluation it is difficult not to overfit the forecast and dressing parameters used to generate probability forecasts; the details of cross-validation can have a large impact. Extending the typical leave-some-out fitting protocol (Hastie et al. 2001; Bröcker and Smith 2008) to include the kernel dressing procedure reduces the sample size of the forecast-outcome archive from eight to seven pairs. This 'true leave-some-out' procedure (Smith et al. 2013) will necessarily increase the sampling uncertainty, reflected through bootstrap resampling. In the case of the ENSEMBLES forecasts, adopting a true leave-some-out procedure reduces the apparent significance of the results; failing to introduce such a procedure, however, risks both information contamination and the suggestion that there is more skill than is to be expected in the simulation models. The most appropriate path cannot be determined with confidence until additional data becomes available.

# 5. Results

The skill of each of the four ENSEMBLES decadal prediction models has been evaluated relative to Dynamic Climatology (DC). The HadGem2 forecast distributions are shown as fan charts in figure 4 as an example. These forecast distributions tend to capture the observed global mean temperature, although the verification falls outside the $5^{th}$-$95^{th}$ percentile of the distribution more often than the expected 10% of the time. The distributions from the other ENSEMBLES simulation models (illustrated in the supplementary material) produce similar results.

A set of forecast distributions for the DC model is shown in figure 5. This model was launched every year between 1960 and 2000, although only every fifth launch is illustrated, in keeping with the ENSEMBLES forecast launch dates. The increased number of launches for the DC model, each with a larger ensemble, allows more accurate statistics on its performance over the same range of the available observational data. Forecasts from the DC model show a similar distribution across each forecast launch, unlike those of the ENSEMBLES models. The verification also falls outside the $5^{th}$-$95^{th}$ percentile of the DC distributions on several occasions, similar to the distributions produced for the simulation models.

Figure 6 shows the performance of all four ENSEMBLES simulation models and the DC empirical model in terms of Ignorance as a function of lead time. To test whether one model is systematically better than another requires considering the relative performance directly. The Ignorance of each model is computed relative to the static climatology shown in figure 7. True leave-some-out cross-validation is applied throughout. When the relative Ignorance is less than zero the model has skill relative to the static climatology. If the bootstrap resampling intervals of a model overlap zero, the model may be less skillful than the static climatology. In fact none of the simulation models consistently outperform the DC empirical model, which has among the lowest Ignorance scores. Figure 6 shows that the DC model significantly outperforms the static climatology across all lead times, on average placing approximately twice the probability mass on the verification ($S_{rel} \approx -1.0$). The DC model using only launch dates every fifth year (to introduce a sampling uncertainty comparable to those of the ENSEMBLES model forecasts) shows a similar result but with slightly larger bootstrap resampling intervals as expected. For each of the ENSEMBLES models variations in skill between forecasts (for a given lead time) prevent the establishment of significant skill relative to the static climatology, despite the fact that both the IFS/HOPE and ARPEGE4/OPA models consistently produce relative Ignorance scores below zero at most lead times. The HadGem2 and ARPEGE4/OPA models, however, indicate that significant skill relative to static climatology can be established for early lead times. It is no surprise that the DC model performs better than the static climatology, since an increase in skill is almost certain to come from initialising each forecast to the observed temperature value at the forecast launch.

Figure 8 shows the performance of each of the models relative to forecasts of persistence. Once again the DC model consistently shows relative Ignorance scores below zero across most lead times, while the ARPEGE4/OPA model scores below zero for early lead times (up to a lead time of five years), suggesting that forecasts from these models are more skillful than a persistence forecast over this range. In both cases the resampling bars cross the zero relative skill axis, clouding the significance of the result.

The skill of the ENSEMBLES simulation model forecasts is illustrated relative to the DC model in figure 9. None of the models in the ENSEMBLES multi-model ensemble demonstrates significant skill above the DC model at any lead time for global mean temperature. In fact all four simulation models show systematically less skill than the DC model. Similar results are found at smaller spatial scales (specifically the Giorgi regions (Giorgi 2002)), where the DC empirical model tends to outperform each of the ENSEMBLES simulation models (see the supplementary material).

The ECHAM5 model generally has the least skill out of the ENSEMBLES models, particularly for global mean temperature, with DC outperforming this model by several bits at lead times of up to ten years, although the bootstrap resampling intervals often overlap the zero line and also overlap with the intervals from the other simulation models in figure 9. At global mean temperature scales the ARPEGE4/OPA model tends to perform better than the other ENSEMBLES models, perhaps surprisingly, since the raw simulation hindcasts from ARPEGE4/OPA contain a particularly large (but consistent) model drift relative to the other simulation models. Models requiring empirical drift corrections are less likely to produce realistic forecasts in a changing climate than they are in the current climate. Over the smaller spatial scales considered (the Giorgi regions) the ARPEGE4/OPA model no longer outperforms the other simulation models; no one ENSEMBLES model emerges as significantly better than any other (see Supplementary Material).

The poor performance of the ECHAM5 simulation model might at first appear as a surprise, since the ensemble members from this model appear to be relatively close to the target values in figure 1. Note, however, that ECHAM5 initialises (and thus forecasts) model anomalies, not physical temperatures; the model forecasts then yield forecast model anomalies. In this case then, the systematic error of the model is partially accounted for when the model forecast anomalies are translated back into physical temperatures. The offset applied within the kernel dressing procedure levels the playing field by accounting for the systematic errors in the other simulation models; the figures indicate that while ECHAM5 may suffer less model drift due to this process (Keenlyide et al. 2005) it does not produce more skillful probability forecasts than the other ENSEMBLES simulation models.

The ENSEMBLES experimental design also contains a perturbed physics ensemble from the UK Met Office Decadal Prediction System (DePreSys) (Doblas-Reyes et al. 2010), in which nine perturbed physics ensemble members are considered over the same set of hindcast launch dates. The DePreSys simulations contain only one initial condition ensemble member for each model version. In this case, the offset and kernel parameters must be determined for each model version separately and the lack of any information on sensitivity to initial conditions limits the practical evaluation of the perturbed physics ensemble. The DePreSys hindcasts are therefore not considered for analysis here.

While hindcast experiments can never provide true "out of sample" evaluation of a forecasting system, it is possible to deny empirical models access to data observed after each launch date. In addition to the denial of what were effectively future observations, it is also necessary to illustrate that the skill of these *Prelaunch* empirical models[8] does not depend sensitively on parameter tuning, as it is implausible that such tuning could have been done in real-time. The results reported below are robust to variations in the free parameters in the Prelaunch DC model (see Supplementary Material).

Two Prelaunch empirical models were considered. The first is simply a direct climatology model where the observation archive is restricted to values prior to each launch date. The results are similar, in fact sometimes slightly better than, the standard DC model. Figure 10 shows the skill of the Prelaunch DC model with a kernel width of ($\sigma = 0.08$ and $\sigma = 0.02$) relative to the standard DC model, constructed under cross-validation; performance is robust

[8] Arguably our "Prelaunch" model could be called a "simulated real-time" model. we resist this inasmuch as the "future" was known when the experiment was designed, even though only the prelaunch observations were used in constructing the model. "Prelauch" should be read to imply only that the data used was restricted to that dated before the forecast launch date, it does not imply that (the impact of) all information gleaned since that date was somehow forgotten.

to decreasing this width by more than an order of magnitude. A Prelaunch Trend model was also constructed to determine if the observed skill was due to a linear trend. The Prelaunch Trend model simply extends the linear fit to the observations from a fixed start-date (say, 1950) to the launch date, and then uses the standard deviation of the residuals as the kernel width. The Prelaunch DC is more skillful than the Prelaunch Trend model, as shown in figure 10. This result is robust to changing the start-date back towards 1900 (see Supplementary Material). It is important to stress that this trend model is not being advocated as a candidate empirical model, but only to address the specific question of whether the skill of the DC model comes only from the observed trend in global mean temperature. Much more effective methods for estimating statistical time-series models are available in this context (see for example (Fildes and Kourentzes 2011)).

The results presented highlight several features for the experimental design of ensemble prediction systems and the impact that design has for the evaluation of probabilistic forecasts. In hindcast experiment design, the number and type of ensemble members considered not only impact on the resolution of the prediction system, but also on the quality of the evaluation methodology: in the kernel dressing approach this impacts the accuracy of the estimated kernel offset and spread parameters, as well as the cross-validation procedure. Sample size plays a major role and has consequences for the design of experiments and their evaluation. In particular the number of available forecasts and ensemble members can heavily influence the significance of the results, especially when the forecast-outcome archive is small. Large initial condition ensembles more clearly distinguish systematic model drift at a particular initial state from sensitivity to small changes in that initial state. Singleton ensembles, as in DePreSys, do not allow such a separation. With only a relatively short forecast-outcome archive and a small number of ensemble members per hindcast launch, the evaluation of the probabilistic forecasts suffers from large sampling uncertainties. While it may not be possible to extend the duration of the observations, increasing the ensemble size can resolve some of the ambiguities involved in the cross-validation stage. In the case of

DePreSys, it is suggested that future perturbed physics hindcast designs would benefit from including initial condition perturbations, as well as different model versions. Further improvements, in terms of increasing the statistical significance of the probabilistic evaluation, may be made by extending the size of the forecast-outcome archive further into the past, or where this is not possible, including intermediate launch dates to increase the sample size for the purpose of fitting the kernel dressing parameters.

# 6. Conclusions



The quality of decadal probability forecasts from the ENSEMBLES simulation models has been compared with that of reference forecasts from several empirical models. In general, the Stream 2 ENSEMBLES simulation models demonstrate less skill than the empirical DC model across the range of lead times from one to ten years. The result holds for a variety of proper scoring rules including Ignorance (Good 1952), the Proper Linear Score (PL) (Jolliffe and Stephenson 2003) and the continuous ranked probability score (CRPS) (Bröcker and Smith 2006). A similar result holds on smaller spatial scales for the Giorgi Regions (see Supplementary Material). These new results for probability forecasts are consistent with evaluations of root-mean-square errors of decadal simulation models with other reference point forecasts (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012; Weisheimer et al. 2009). The DC probability forecasts often place up to 4 bits more information (or $2^4$ times more probability mass) on the observed outcome than the ENSEMBLES simulation models.

In the context of climate services, the comparable skill of simulation models and empirical models suggests that the empirical models will be of value for blending with simulation model ensembles; this is already done in ensemble forecasts for the medium range and on seasonal lead times. It also calls into question the extent to which current simulation models successfully capture the physics required for realistic simulation of the Earth System, and can thereby be expected to provide robust, reliable predictions (and, of course, to outperform empirical models) on longer time scales.

The evaluation and comparison of decadal forecasts will always be hindered by the relatively small samples involved when contrasted with the case of weather forecasts; the decadal forecast-outcome archive currently considered is only half a century in duration. Advances both in modelling and in observation, as well as changes in the Earth's climate, are likely to mean the relevant forecast-outcome archive will remain small. One improvement that could be made to clarify the skill of the simulation models is to improve the experimental design of hindcasts, in particular to increase the ensemble size used. For the ENSEMBLES models, each simulation ensemble consisted of only three members launched at five year intervals. Larger ensembles and more frequent forecast launch dates can ease the evaluation of skill without waiting for the forecast-outcome archive to grow larger[9].

The analysis of hindcasts can never be interpreted as an "out of sample" evaluation. The mathematical structure of simulation models, as well as parameterizations and parameter values, have been developed with knowledge of the historical data. Empirical models with a simple mathematical structure suffer less from this effect. Prelaunch empirical models based on the DC structure and using only observations before the forecast launch date also outperform the ENSEMBLES simulation models. This result is robust over a range of ensemble interpretation parameters (that is, variations in the kernel width used). Both Prelaunch Trend models and persistence models are less skillful than the DC models considered.

The comparison of near-term climate probability forecasts from Earth Simulation Models with those from Dynamic Climatology empirical models provides a useful benchmark as the simulation models improve in the future. The blending (Bröcker and Smith 2008) of simulation models and empirical models is likely to provide more skillful probability forecasts in

[9]As noted by a reviewer, it is possible that a DC model effectively captures all the available forecast information given the uncertainty in the observations. This suggestion would be supported if the ENSEMBLES models were shown to be able to shadow (Smith 1997) over decades and, even with improved data assimilation and using large ensembles, did not outperform empirical models; on the other hand it could be easily falsified by a single simulation model which convincingly outperformed the empirical models.

Climate Services, both for policy and adaptation decisions. In addition, clear communication of the (limited) expectations for skillful decadal forecasts can avoid casting doubt on well-founded physical understanding of the radiative response to increasing carbon dioxide concentration in the Earth's atmosphere. Finally, these comparisons cast a sharp light on distinguishing whether current limitations in estimating the skill of a model arise from external factors like the size of the forecast-outcome archive, or from the experimental design. Such insights are a valuable product of ENSEMBLES and will contribute to the experimental design of future ensemble decadal prediction systems.

# Acknowledgments

This research was funded as part of the NERC EQUIP project (NE/H003479/1); it was also supported by the EU Framework 6 ENSEMBLES project (GOCE-CT-2003-505539-ENSEMBLES) and by both by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. L.A.S. gratefully acknowledges support of Pembroke College, Oxford. We also acknowledge the helpful comments and insights from Roman Binter, Hailiang Du, Ana Lopez, Falk Niehörster, David Stainforth and Erica Thompson, which helped shape this work, as well as discussions with H. van den Dool, G.-J. van Oldenborgh, A. Weisheimer and two anonymous reviewers which improved an earlier manuscript.



# Appendix: The Stream 2 ENSEMBLES decadal hindcast experiments

The set of decadal hindcast experiments from Stream 2 of the ENSEMBLES project simulations (Doblas-Reyes et al. 2010) have a similar experimental design to the seasonal hindcast experiments discussed in (Weisheimer et al. 2009). The decadal hindcasts consist of a set of initial condition ensembles, containing three ensemble members, initialised at launch, from four forecast systems - ARPEGE4/OPA (CERFACS), IFS/HOPE (ECMWF), HadGem2 (UKMO) and ECHAM5 (IFM-GEOMAR) - to produce a multi-model ensemble. A perturbed physics ensemble containing nine ensemble members from the DePreSys forecast system (based on the HadCM3 climate model) for both initialised and unassimilated simulations also forms part of the ENSEMBLES project. The hindcasts span the period 1960-2005, with simulations from each model launched at 5-year intervals, starting in November of the launch year and run over 10-year integrations. A full initialisation strategy was employed for the atmosphere and ocean using realistic estimates of their observed states (except for ECHAM5, which employed an anomaly initialisation scheme), with all the main radiative forcings prescribed and perturbations of the wind stress and SST fields made to sample initial condition uncertainty of the multi-model ensemble.



# REFERENCES

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Deque, P. Rogel, and A. Weisheimer, 2011: Evaluation of probabilistic quality and value of the ensembles multimodel seasonal forecasts: Comparison with demeter. *Monthly Weather Review*, **139**, 581–607.

Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bulletin of the American Meteorological Society*, **80 (7)**.

Barnston, A. G., et al., 1994: Long-lead seasonal forecasts - where do we stand? *Bulletin of the American Meteorological Society*, **75 (11)**, 2097–2114.

Binter, R., 2011: Applied probabilistic forecasting. Ph.D. thesis, London School of Econmics and Political Science.

Bröcker, J. and L. A. Smith, 2006: Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, **22**, 382–388.

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

Colman, A. and M. Davey, 2003: Statistical prediction of global sea-surface temperature anomalies. *International Journal of Climatology*, **23 (956)**, 1677–1697.

Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer, 2011: Decadal climate prediction with the european centre for medium-range weather forecasts coupled forecast system: Impact of ocean observations. *Journal of Geophysical Research - Atmospheres*, **116 (D19111)**.

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**.

Fildes, R. and N. Kourentzes, 2011: Validation and forecasting accuracy in models of climate change. *International Journal of Forecasting*, **27 (4)**.

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.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. part i: Basic concept. *Tellus*, **A57**, 219–233.

Hagedorn, R. and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, **16 (2)**, 143–155.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The elements of statistical learning*. Springer, New York.

Hawkins, E., J. Robson, R. Sutton, D. Smith, and N. Keenlyside, 2011: Evaluating the potential for statistical decadal predictions of sea surface temperature with a perfect model approach. *Climate Dynamics*, 1–15.

Hewitt, C. D. and D. J. Griggs, 2004: Ensembles-based predictions of climate and their impacts. *Eos, Transactions American Geophysical Union*, **85 (52)**, 566.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Statistical Science*, **14 (4)**.

IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [S. Solomon and D. Qin and M. Manning and Z. Chen and M. Marquis and K. B. Averyt and M. Tignor and H. L. Miller (eds.)]*. 996 pp, Cambridge University Press, Cambridge, United Kingdom and New York, NY USA.

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

Keenlyide, N. S., M. Latif, M. Botzet, J. Jungclaus, and U. Schulzweida, 2005: A coupled method for initializing el nino southern oscillation forecasts using sea surface temperature. *Tellus*, **57A**, 340–356.

Keenlyide, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the north atlantic sector. *Nature*, **453 (06921)**, 84–88.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model cmip5 decadal hindcasts. *Geophysical Research Letters*, **39 (L10701)**.

Krueger, O. and J.-S. von Storch, 2011: A simple empirical model for decadal prediction. *Journal of Climate*, **24**, 1276–1283.

Laepple, T., S. Jewson, and K. Coughlin, 2008: Interannual temperature predictions using the cmip3 multi-model ensemble mean. *Geophysical Research Letters*, **35 (L10701)**.

Lee, T. C. K., F. W. Zwiers, X. Zhang, and M. Tsao, 2006: Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. *Journal of Climate*, **19 (20)**, 5305–5318.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20 (2)**, 130–141.

Meehl, G. A., et al., 2009: Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, **90**, 1467–1485.

Paparella, F., A. Provenzale, L. A. Smith, C. Taricco, and R. Vio, 1997: Local random analogue prediction of nonlinear processes. *Physics Letters A*, **235 (3)**, 233–240.

Pierce, D. W., T. P. Barnett, R. Tokmakian, A. Semtner, M. Maltrud, J. A. Lysne, and A. Craig, 2004: The acpi project, element 1: Initialising a coupled climate model from observed conditions. *Climatic Change*, **62 (1)**, 13–28.

Pohlmann, H., J. H. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with gecco oceanic synthesis: Effects on the north atlantic. *Journal of Climate*, **22**, 3926–3938.

Reifen, C. and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophysical Research Letters*, **36 (L13704)**.

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

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skilful multi-year predictions of atlantic hurricane frequency. *Nature Geoscience*, **3 (1004)**, 846–849.

Smith, L. A., 1992: Identification and prediction of low-dimensional dynamics. *Physica D*, **58 (1-4)**, 50–76.

Smith, L. A., 1994: Local optimal prediction: Exploiting strangeness and the variation of sensitivity to initial condition. *Philosophical Transactions of the Royal Society A*, **348 (1688)**, 371–381.

Smith, L. A., 1997: The maintenance of uncertainty. *Proc. International School of Physics "Enrico Fermi"*, **Course CXXXIII**, 177–246, Societ'a Italiana di Fisica, Italy.

Smith, L. A., 2002: What might we learn from climate forecasts? *Proceedings of the National Academy of Science*, **4 (99)**, 2487–2492.

Smith, L. A., H. Du, F. Niehörster, and E. B. Suckling, 2013: An evaluation of probabilistic skill from ensemble seasonal forecasts. *submitted to the Quarterly Journal of the Royal Meteorological Society*.

Stockdale, T. N., 1997: Coupled ocean-atmosphere forecasts in the presence of climate drift. *Monthly Weather Review*, **125**, 809–818.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the cmip5 experimental design. *http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf*.

Tennekes, H., A. P. M. Baede, and J. D. Opsteegh, 1987: Forecasting forecast skill. *Proceedings ECMWF Workshop on Predictability, ECMWF, Reading, UK*, 277–302.

Troccoli, A. and T. N. Palmer, 2007: Ensemble decadal predictions from analysed initial conditions. *Philosophical Transactions of the Royal Society A*, **365**, 2179–2191.

UK-MetOffice, 2011: 3-month outlook for uk contingency planning. *http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_user_guidance-150.pdf*.

Unger, D., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Monthly Weather Review*, **137 (2365-2379)**.

van den Dool, H. M., 1994: Long-range weather forecasts through numerical and empirical methods. *Dynamics of Atmospheres and Oceans*, **20 (3)**, 247–270.

van den Dool, H. M., 2007: *Empirical methods in short-term climate prediction*. Oxford University Press.

van Oldenborgh, G. J., M. Balmaseda, L. Ferranti, T. Stockdale, and D. Anderson, 2005: Evaluation of atmospheric fields from the ecmwf seasonal forecasts over a 15-year period. *Journal of Climate*, **18**, 3250–3269.

van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, **38 (7-8)**, 1263–1280.

Weigela, A. P. and N. E. Bowlerb, 2009: Can multi-model combination really enhance prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, **135**, 535–539.

Weisheimer, A., et al., 2009: Ensembles - a new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond demeter in forecasting tropical pacific ssts. *Geophysical Research Letters*, **36 (L21711)**.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences, 3rd Edition*, Vol. 100. Academic Press.

# List of Figures

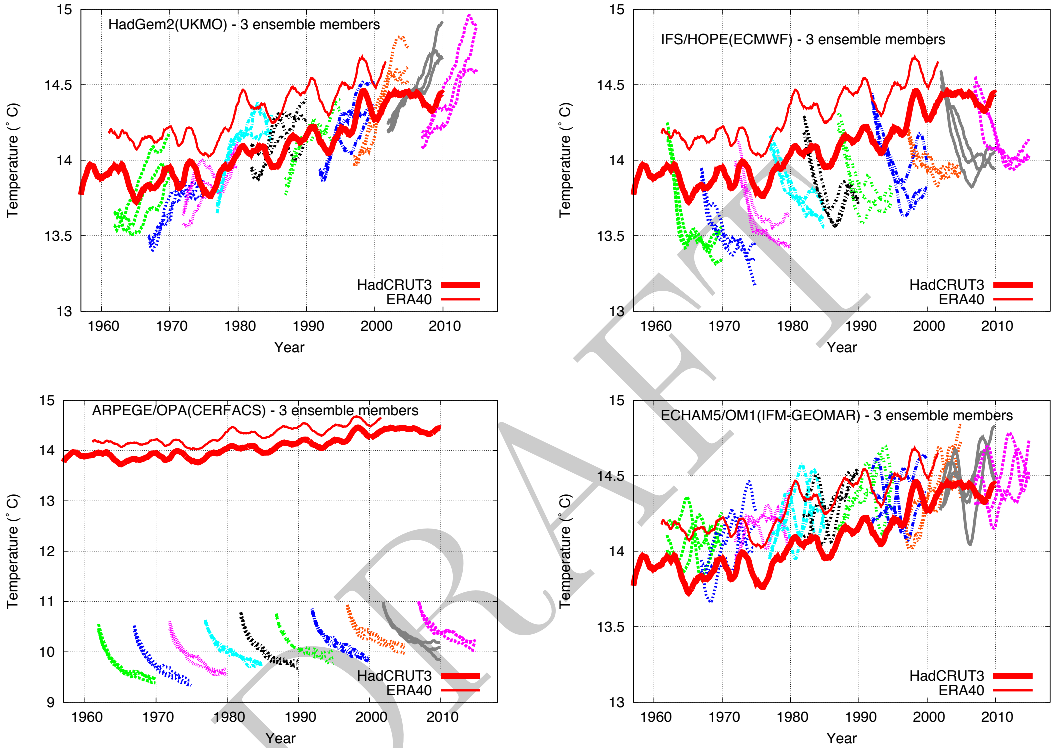


FIG. 1. Global mean temperature (2 year running mean) for the four forecast systems - HadGem2 (UKMO), IFS/HOPE (ECMWF), ARPEGE4/OPA (CERFACS) and ECHAM5 (IFM-GEOMAR) - that form Stream 2 of the ENSEMBLES decadal hindcast simulations (Doblas-Reyes et al. 2010). HadCRUT3 observations and ERA40 reanalysis are also shown for comparison. Note that the scale on the vertical axis for the ARPEGE4/OPA model is different to the other three panels, reflecting the larger bias in this model.

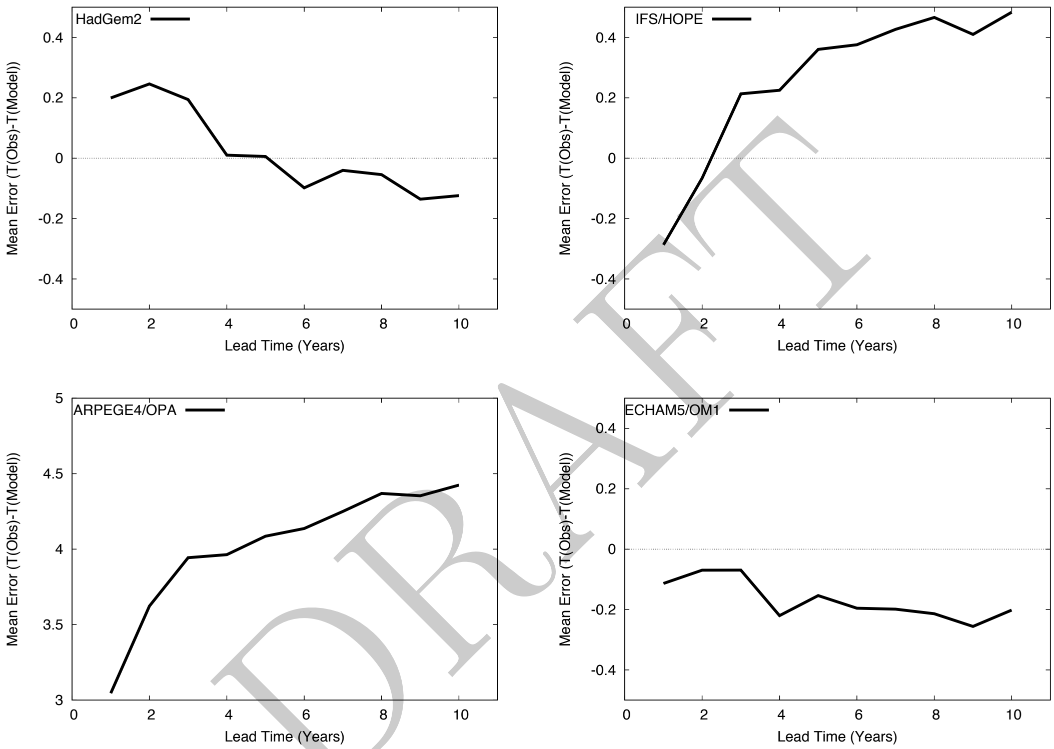




FIG. 2. Mean forecast error as a function of lead time across the set of decadal hindcasts for each of the ENSEMBLES simulation models as labelled. Note that the scale on the vertical axis for the ARPEGE4/OPA model is different to the other three panels, reflecting the larger bias in this model.

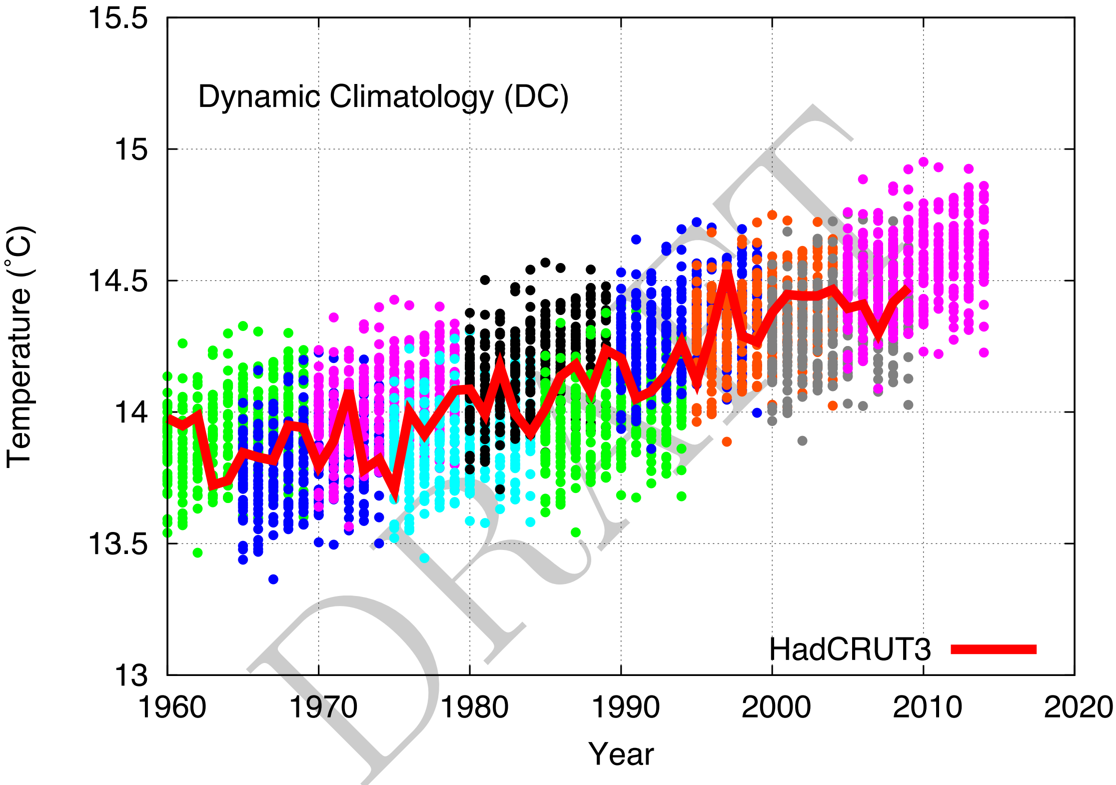


Fig. 3. Dynamic climatology (DC) over the period of the ENSEMBLES hindcasts (figure 1). HadCRUT3 (from which the DC model is constructed) is shown for comparison.

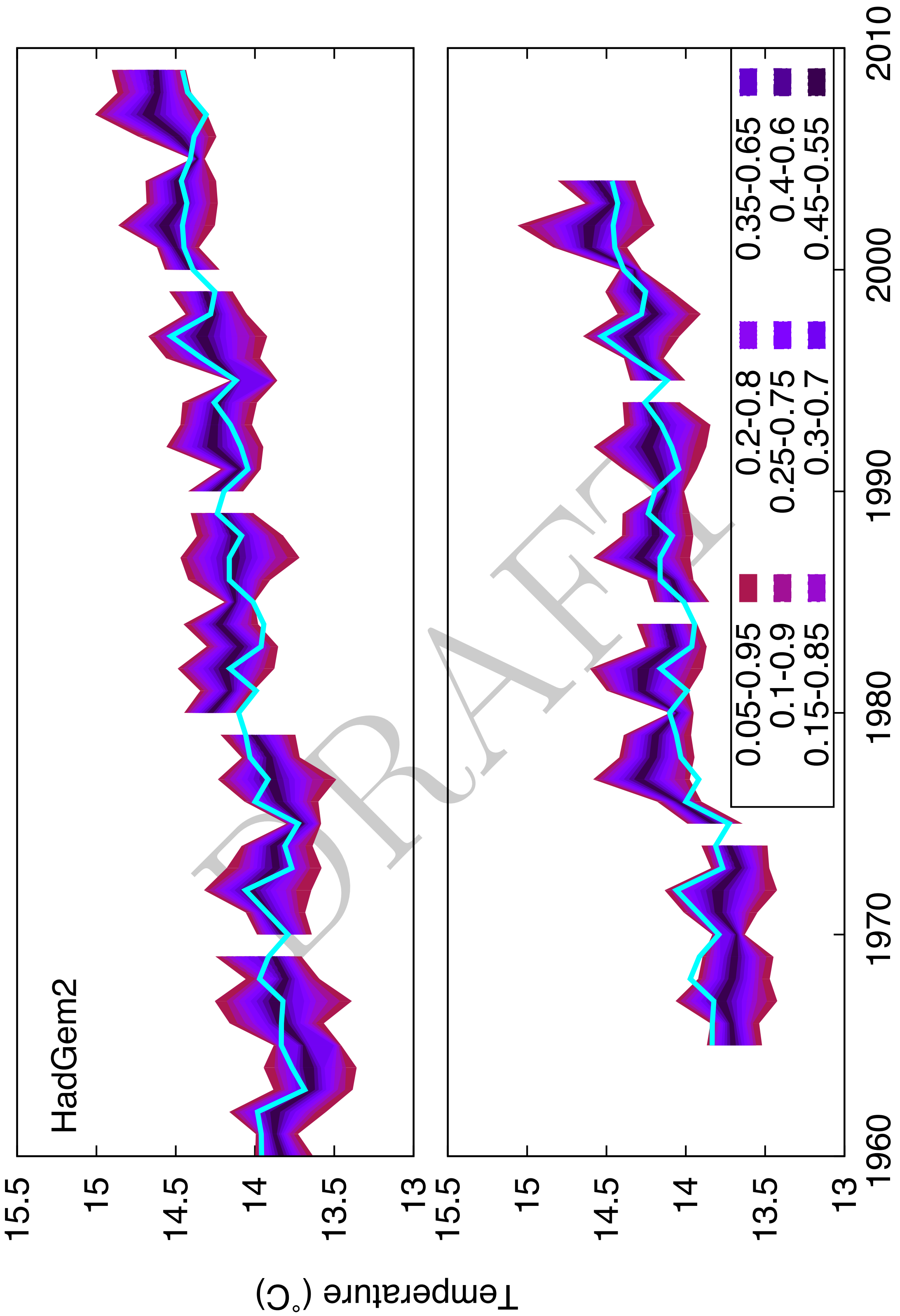




FIG. 4. Forecast distributions for HadGem2 (UKMO) for the 5-95$^{th}$ percentile. The HadCRUT3 observed temperatures are shown in blue. The forecasts are ten years long and lauched every five years, and so the fan charts would overlap; to avoid this they are presented on two panels. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965).

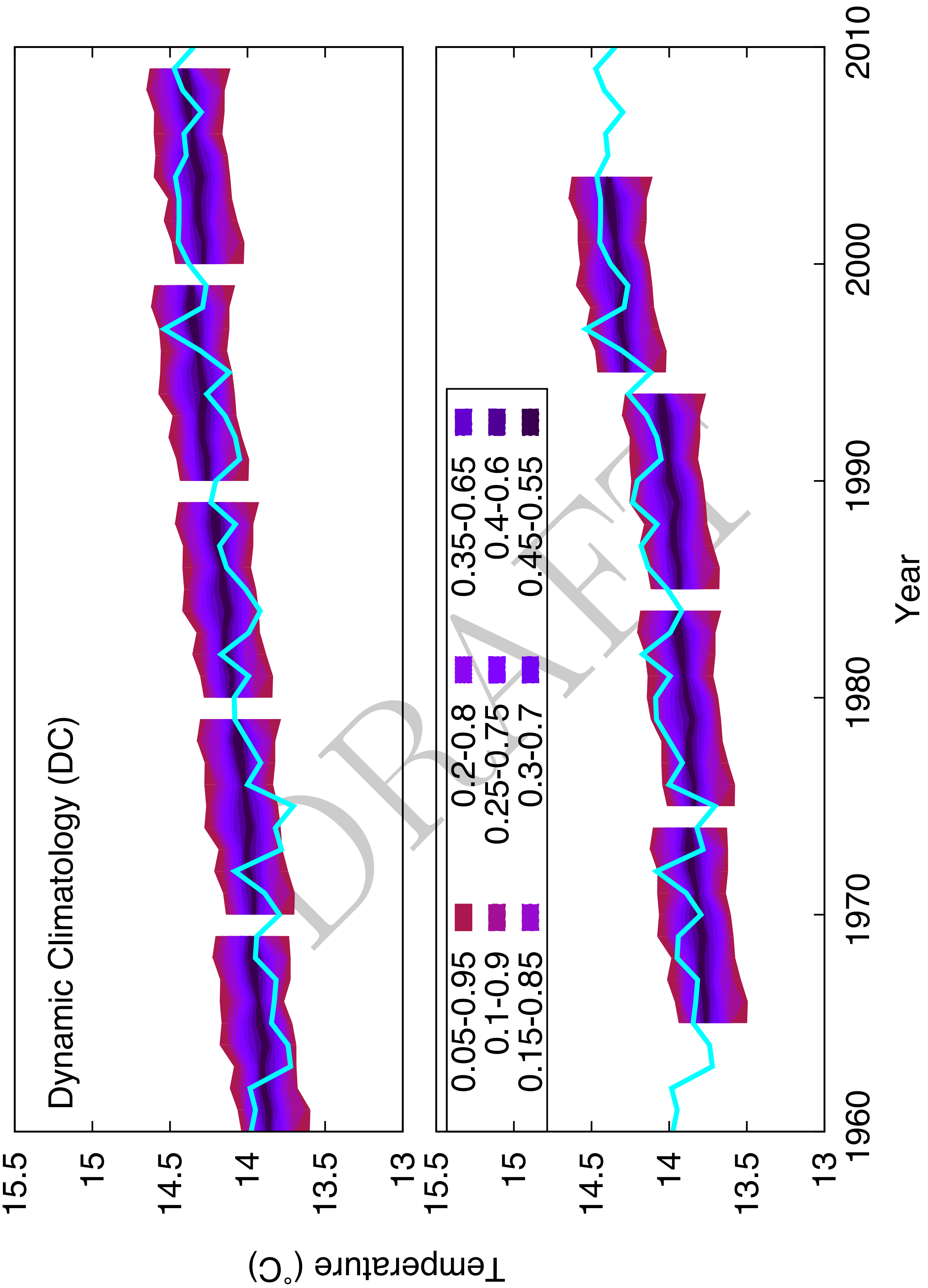


FIG. 5. Forecast distribution for every fifth launch from the Dynamic Climatology (DC) model for the 5-95$^{th}$ percentile. The HadCRUT3 observed temperatures are shown in blue. The forecasts are ten years long and lauched every five years, and so the fan charts would overlap; to avoid this they are presented on two panels. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965).

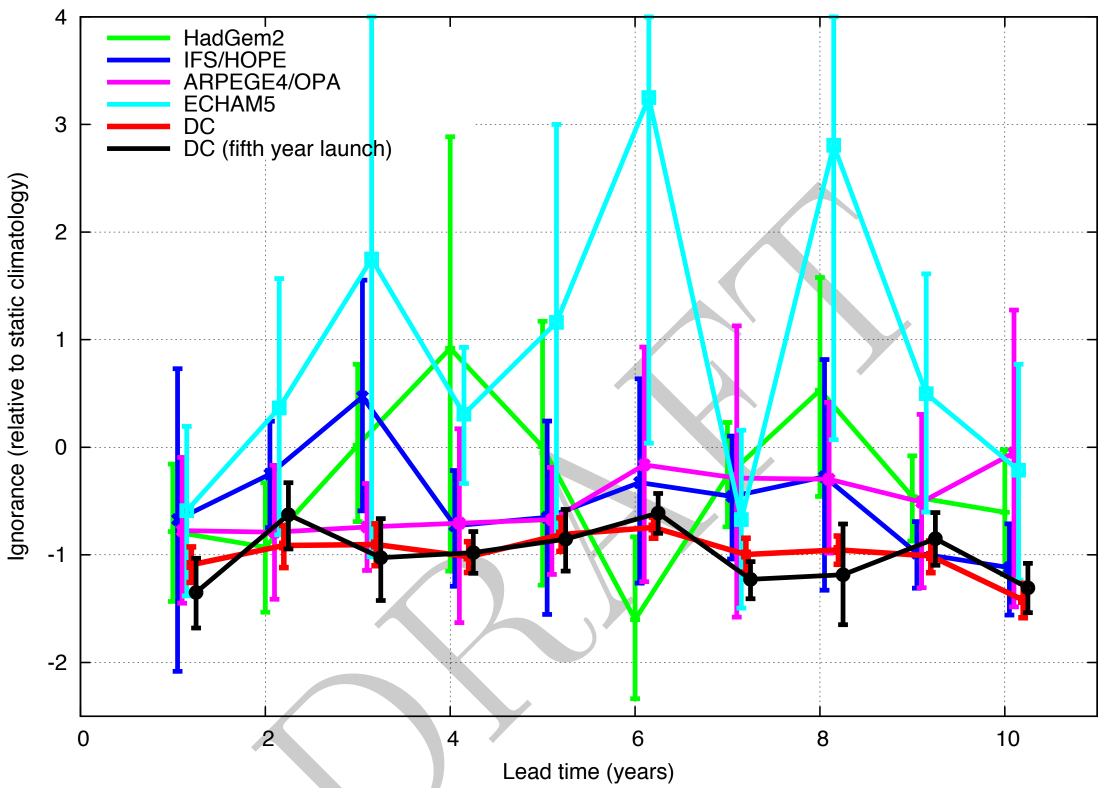


Fig. 6. Ignorance as a function of lead time for each of the four ENSEMBLES hindcast simulation models and the DC model relative to the static climatology. The bootstrap resampling intervals are illustrated at the 10-90$^{th}$ percent level. The DC model is shown to be significantly more skillful than static climatology at all lead times, whereas the ARPEGE4/OPA and IFS/HOPE models are significantly more skillful than static climatology at early lead times.

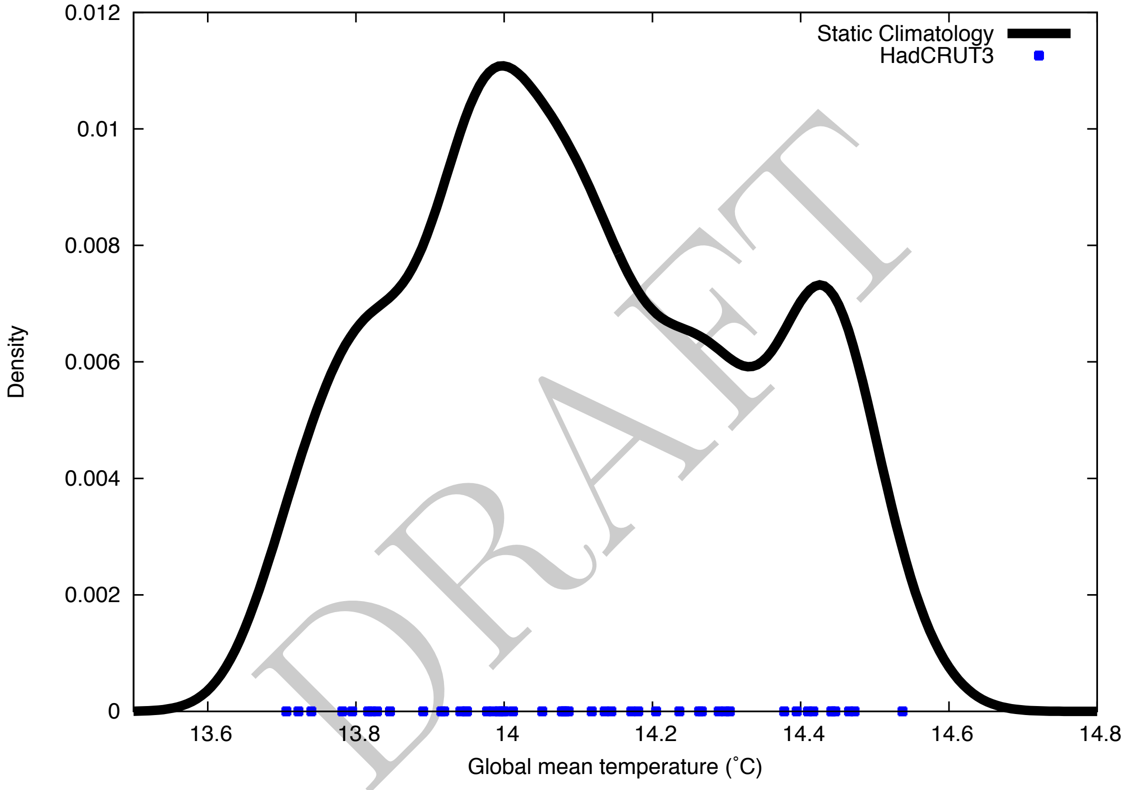


FIG. 7. Probability density for the static climatology used in the paper with observations over the period 1960-2010 (from HadCRUT3) illustrated as points on the x-axis for reference.

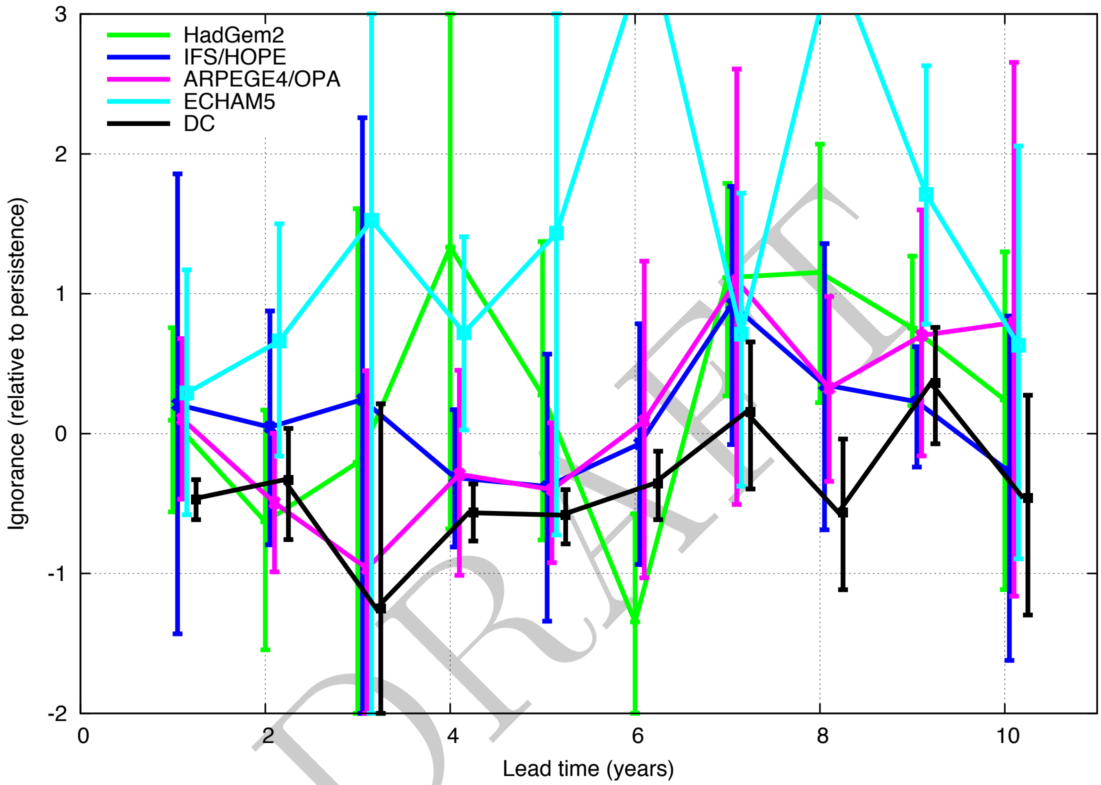


FIG. 8. Ignorance of the ENSEMBLES models and DC relative to persistence forecasts as a function of lead time. The DC model has negative relative Ignorance scores up to 6 years ahead, indicating it is significantly more skillful than persistence forecasts at early lead times. The ENSEMBLES models tend to have positive scores, particularly at longer lead times, with bootstrap resampling intervals that overlap with the zero skill line. The bootstrap resampling intervals are illustrated at the 10-90$^{th}$ percent level.

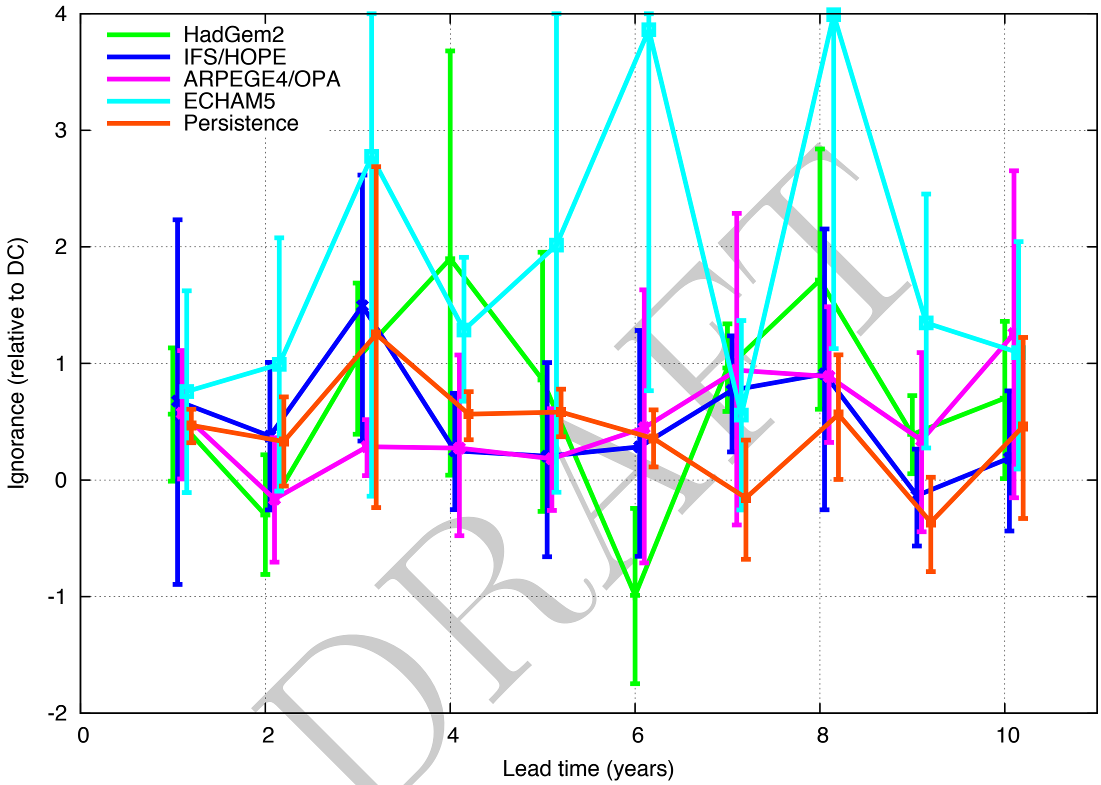


FIG. 9. Ignorance of the ENSEMBLES models relative to DC as a function of lead time. The bootstrap resampling intervals are illustrated at the 10-90$^{th}$ percent level. Note that the simulation models tend to have positive scores (less skill) than the DC model at every lead time.

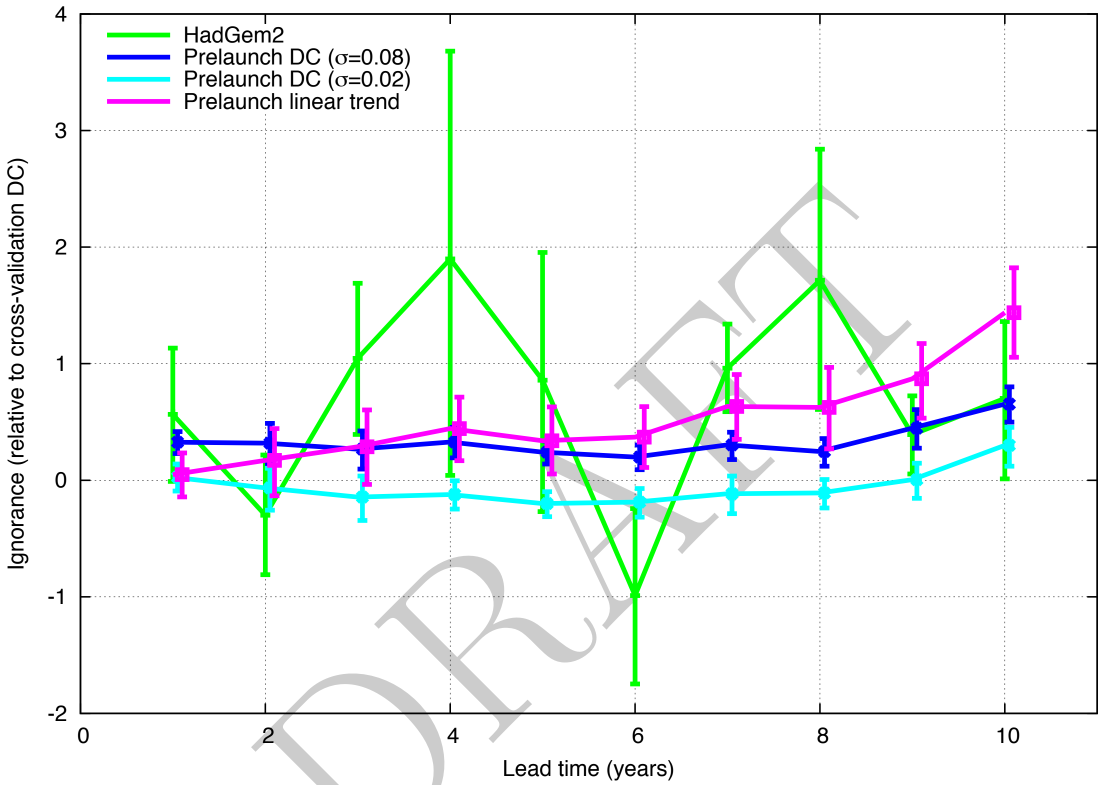


FIG. 10. Ignorance of the Prelaunch DC and Prelaunch Trend models relative to the standard DC model as a function of lead time. The HadGem2 model from ENSEMBLES is also shown. It is shown that the Prelaunch DC model is not significantly less skillful than the standard DC model and is robust to variations in parameter tuning. The Prelaunch linear trend model is, however, generally shown to be less skillful than the standard DC model. The bootstrap resampling intervals are illustrated at the 10-$90^{th}$ percent level.

# 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

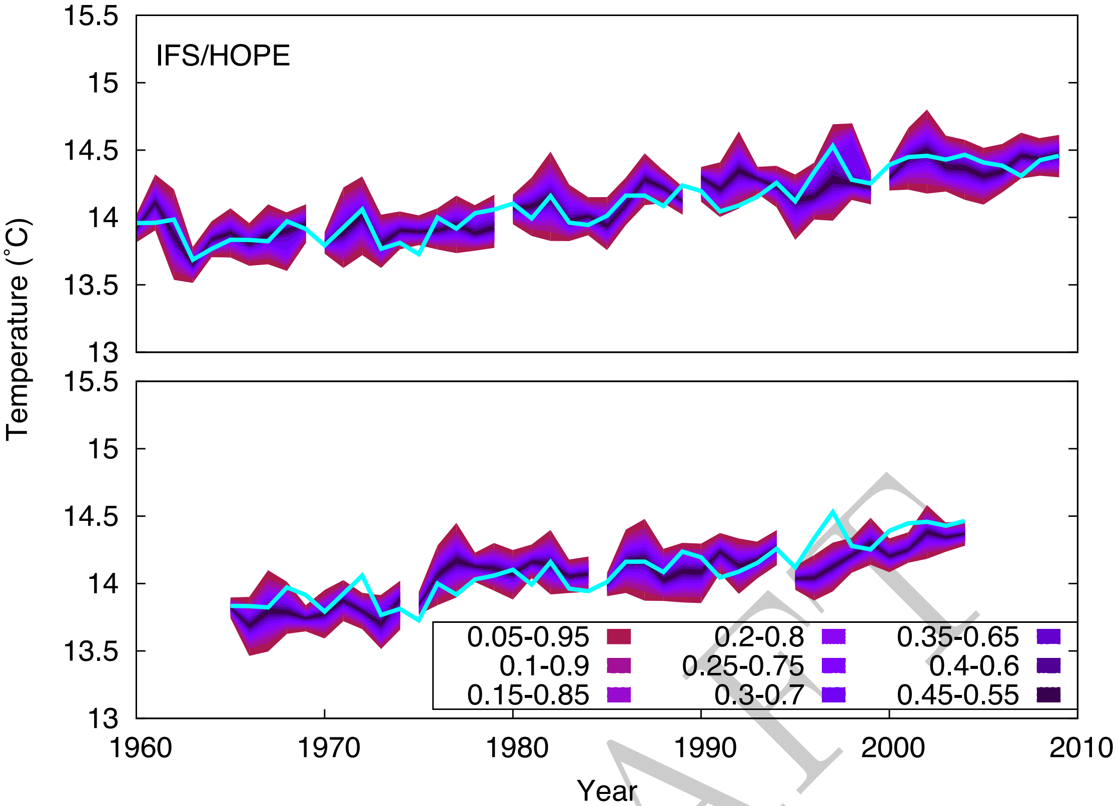




FIG. 1. Forecast distributions for IFS/HOPE (ECMWF) for the 5-95$^{th}$ percentile. The HadCRUT3 observed temperatures are shown in blue. Each forecast is ten years long and they are launched every five years. To avoid overlap of the fan charts they are presented on two panels. The top (bottom) panel 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.

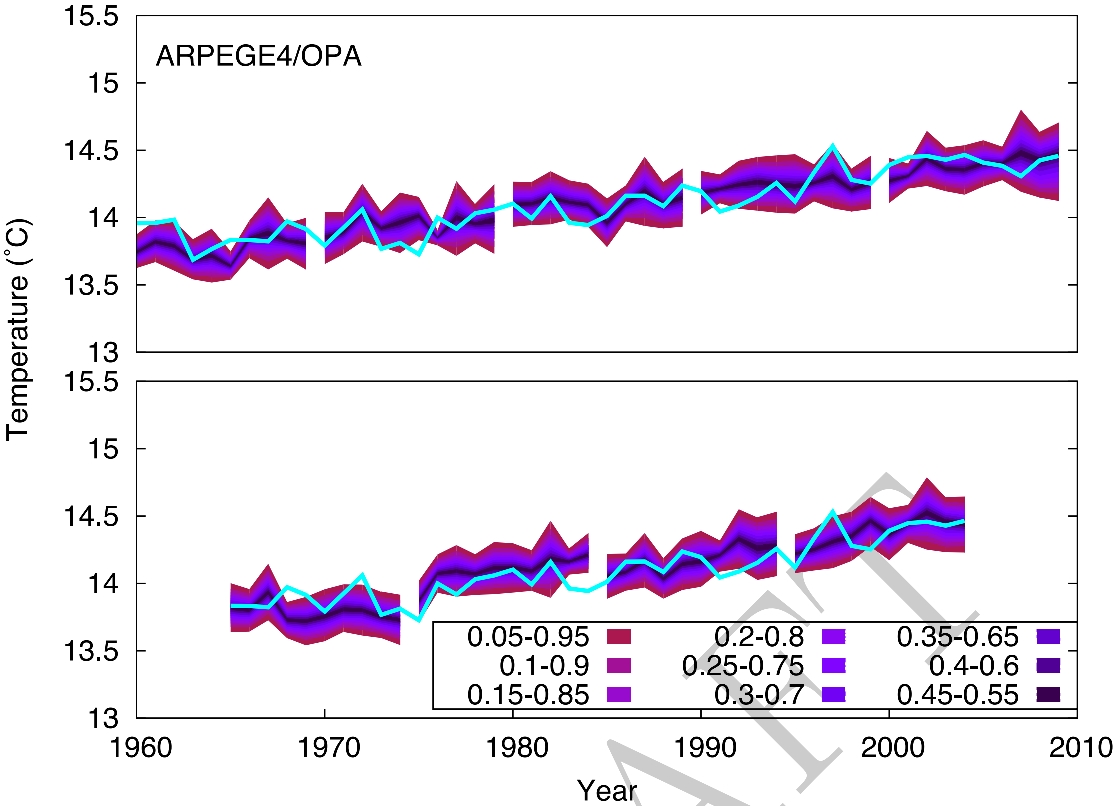


FIG. 2. Forecast distributions for ARPEGE/OPA (CERFACS) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside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

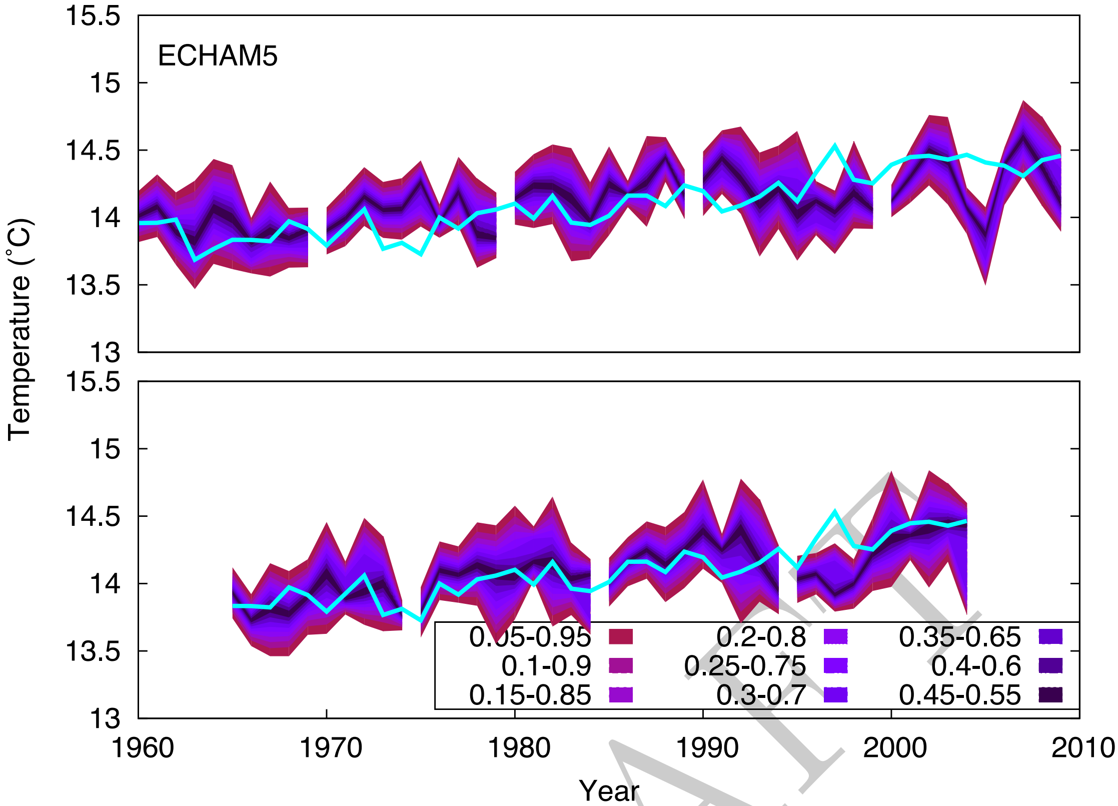


FIG. 3. Forecast distributions for ECHAM5 (IFM-GEOMAR) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature falls outside the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

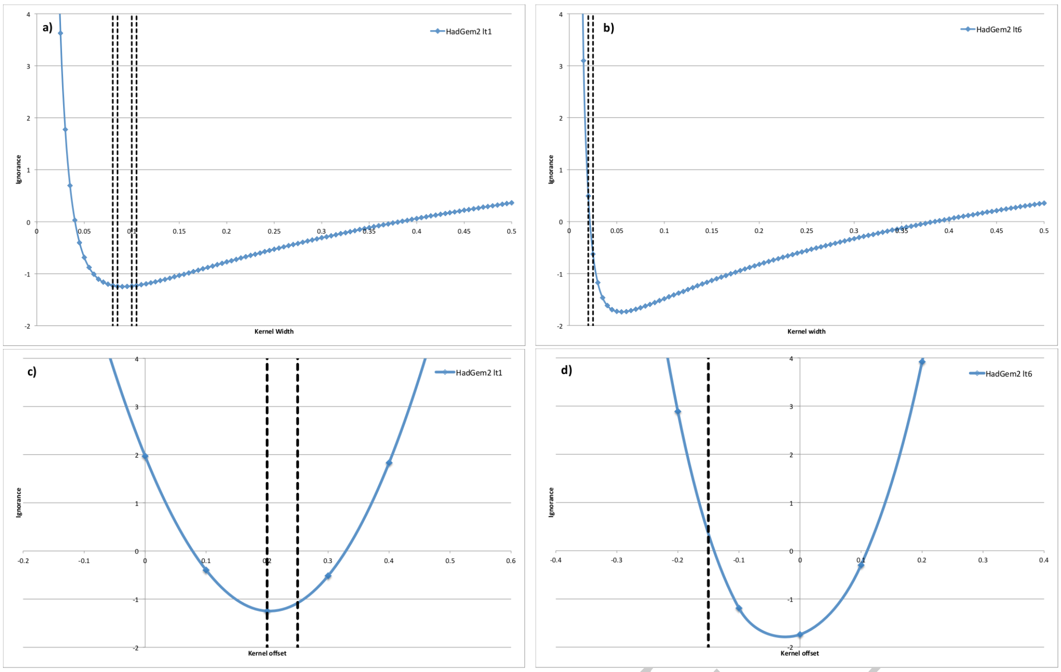




FIG. 4. Ignorance as a function of kernel dressing parameters over the full set of hindcast simulations (*i.e.* with no cross-validation) for the HadGem2 model at lead time one (a and c) and lead time six (b and d). The top panels (a and b) show the score as a function of the kernel width parameter and 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.

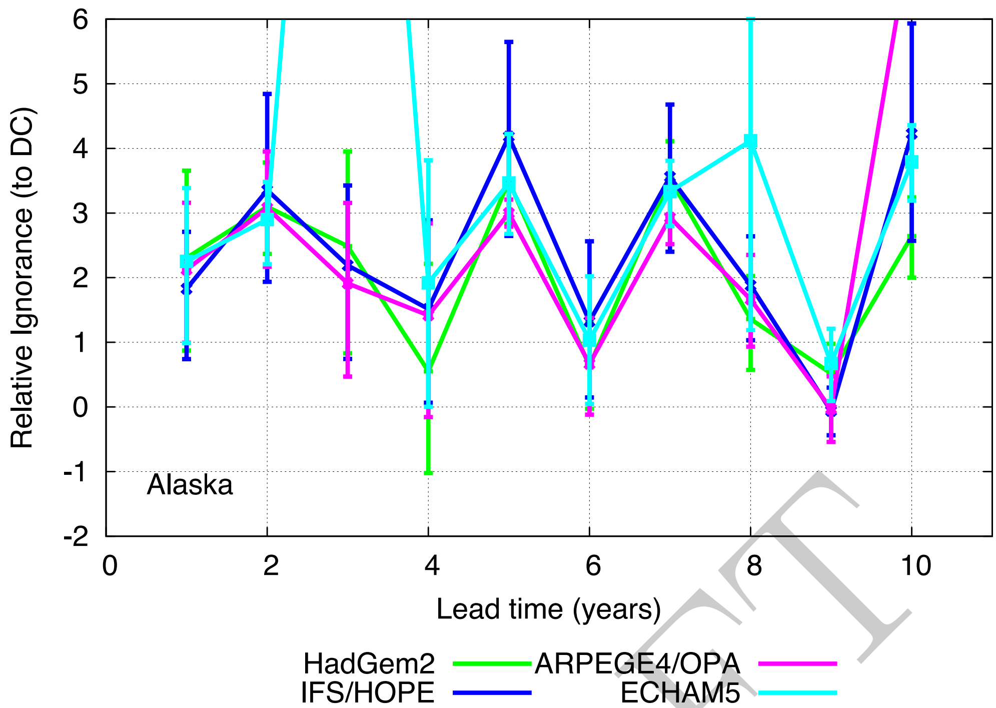


FIG. 5. Ignorance of the ENSEMBLES simulation models relative to the DC model for Alaska. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

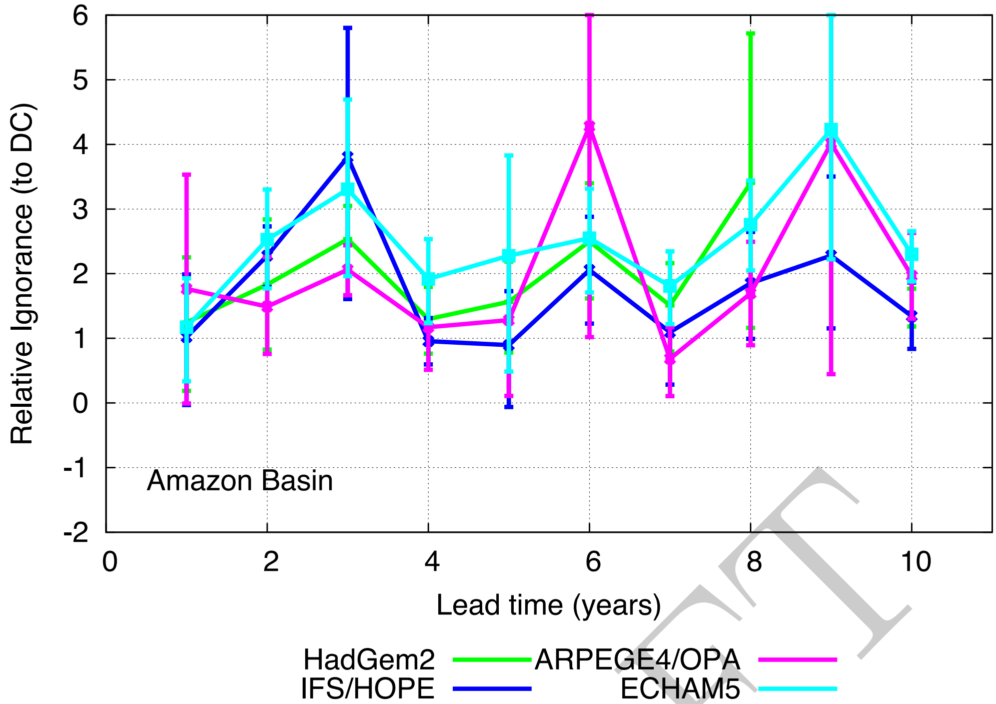


FIG. 6. Ignorance of the ENSEMBLES simulation models relative to the DC model for Amazon Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

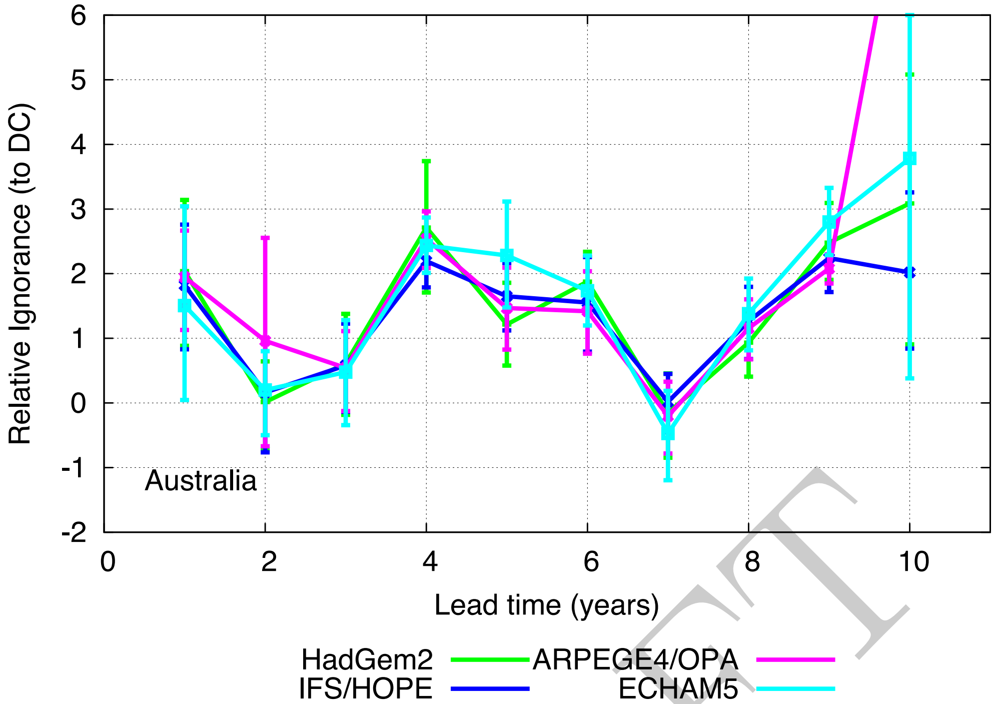


FIG. 7. Ignorance of the ENSEMBLES simulation models relative to the DC model for Australia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

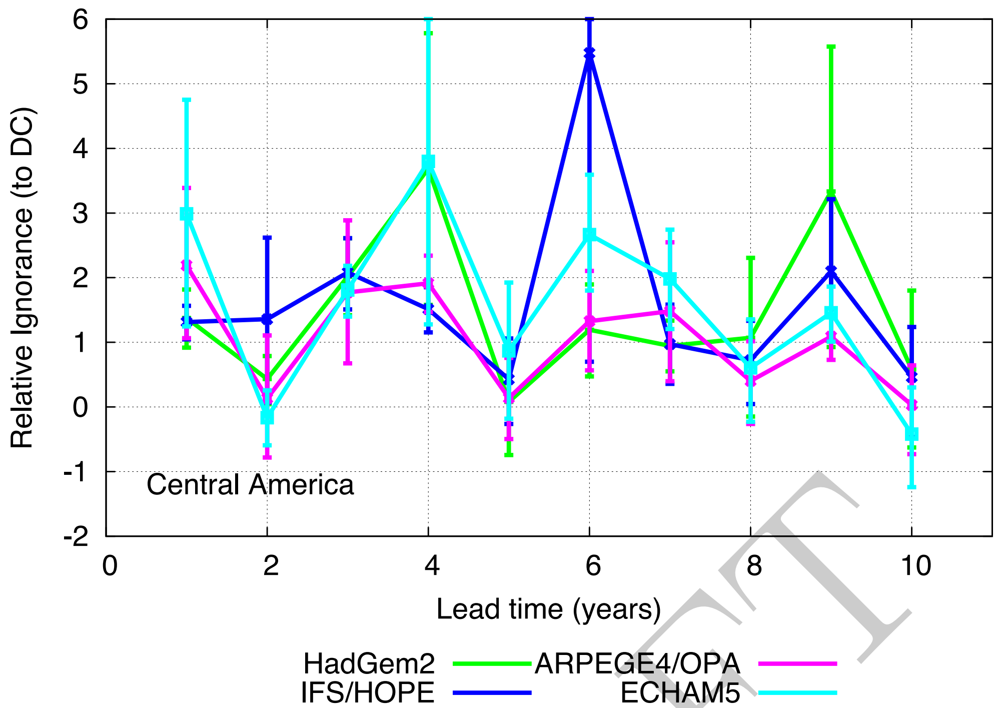


FIG. 8. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

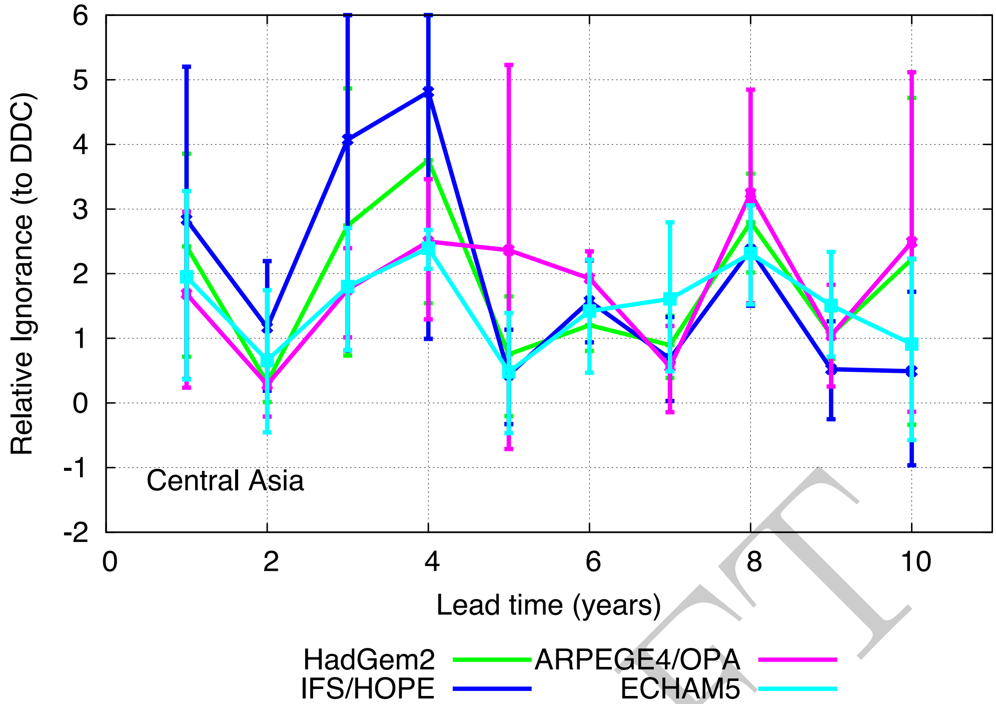


FIG. 9. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

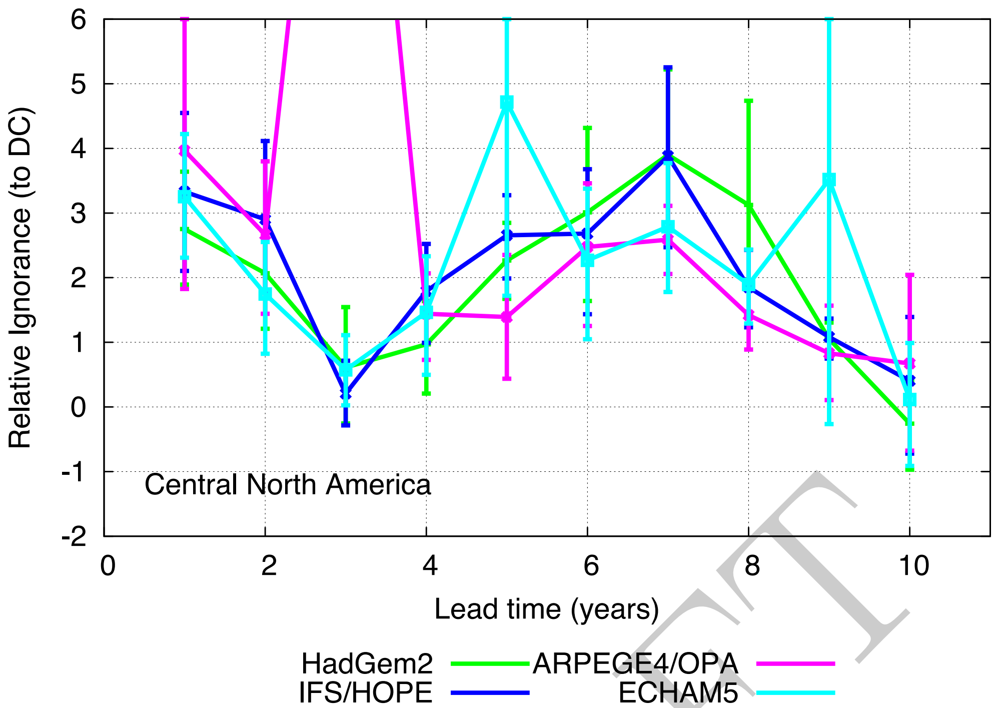


FIG. 10. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

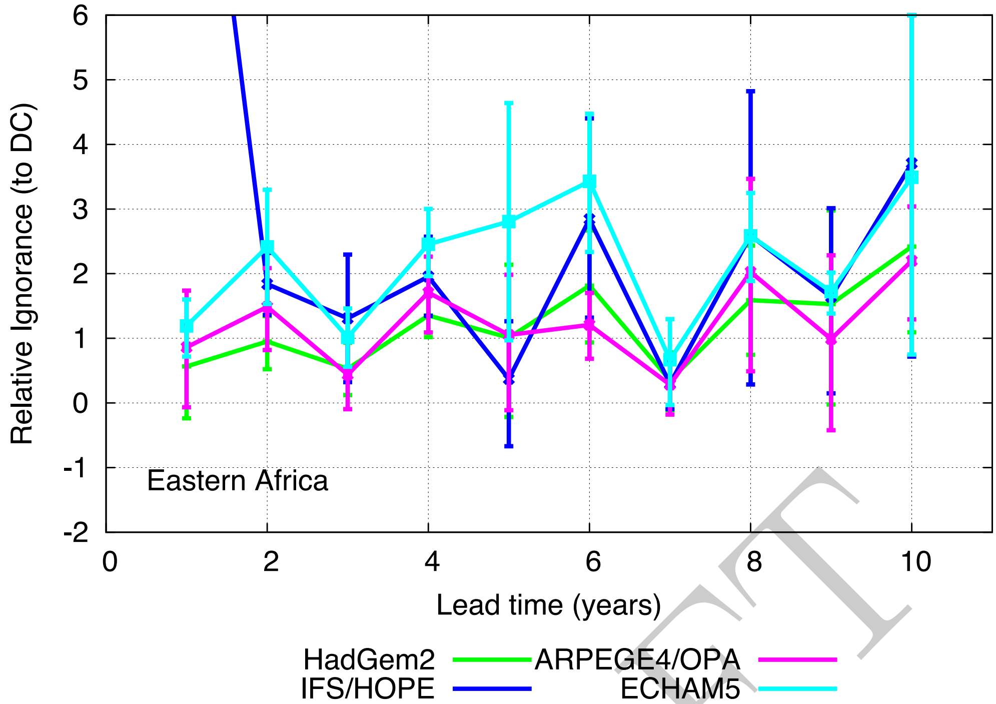


FIG. 11. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

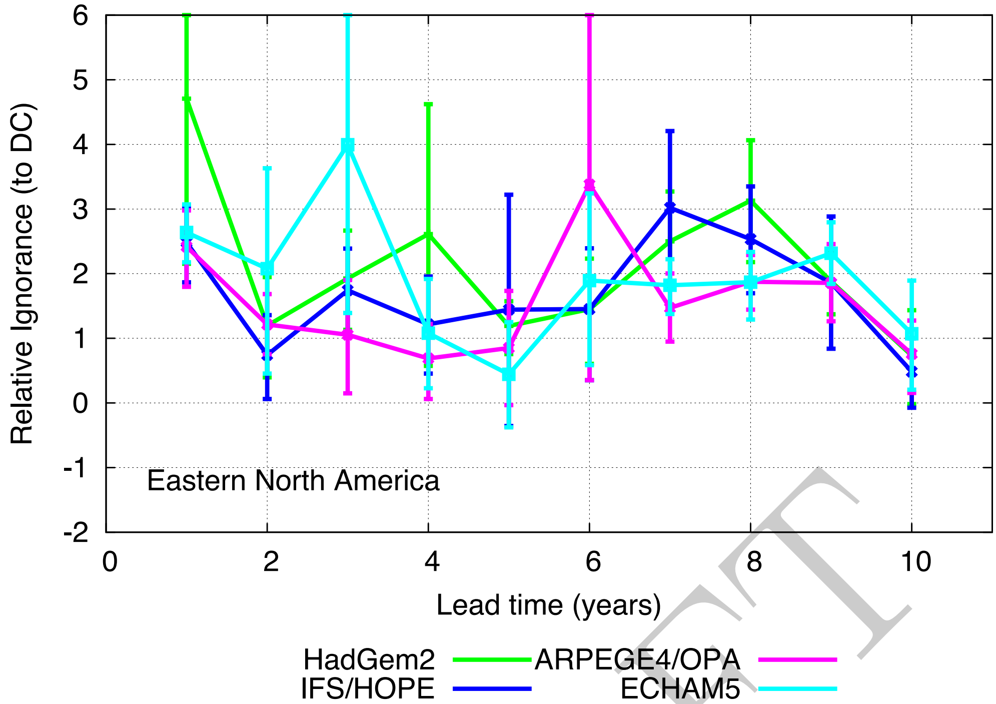


FIG. 12. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

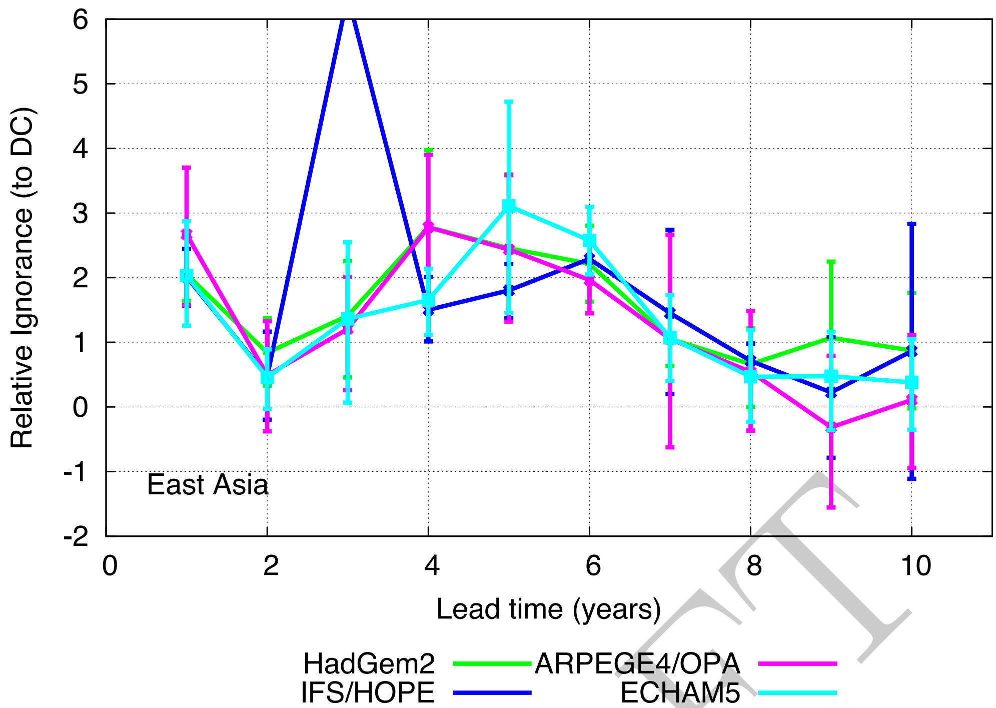


FIG. 13. Ignorance of the ENSEMBLES simulation models relative to the DC model for East Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

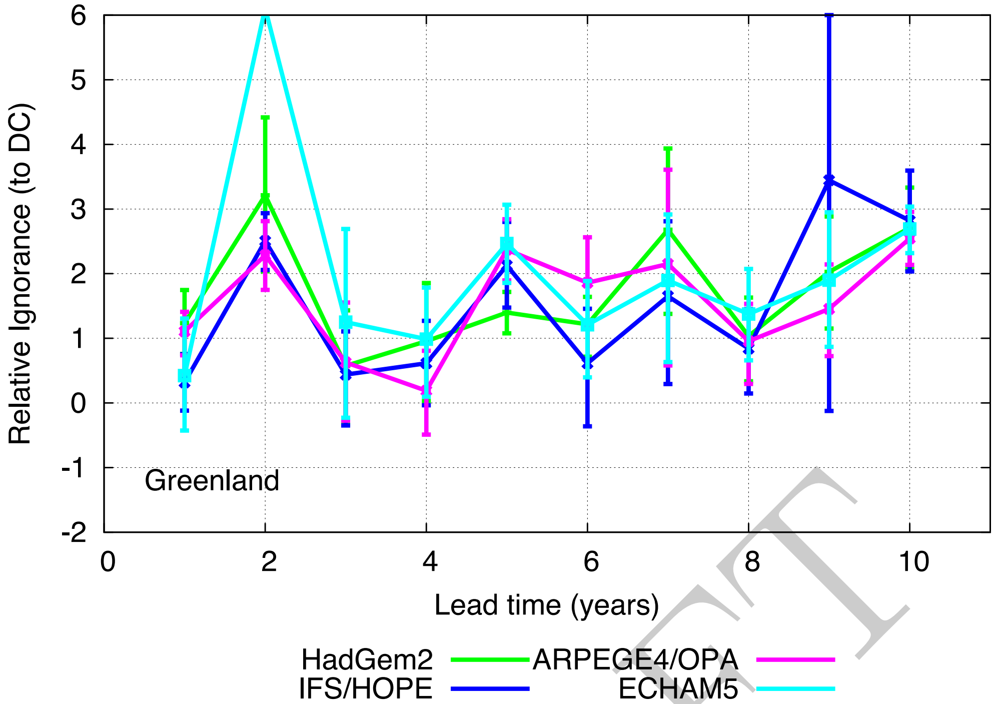


Fig. 14. Ignorance of the ENSEMBLES simulation models relative to the DC model for Greenland. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

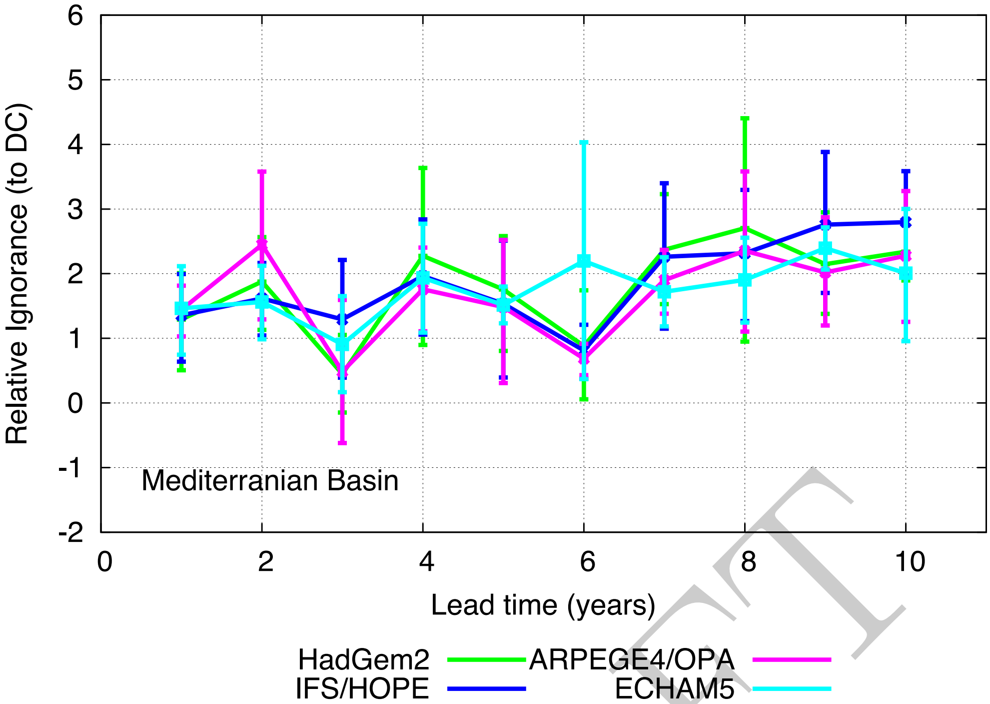


FIG. 15. Ignorance of the ENSEMBLES simulation models relative to the DC model for Mediterranian Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

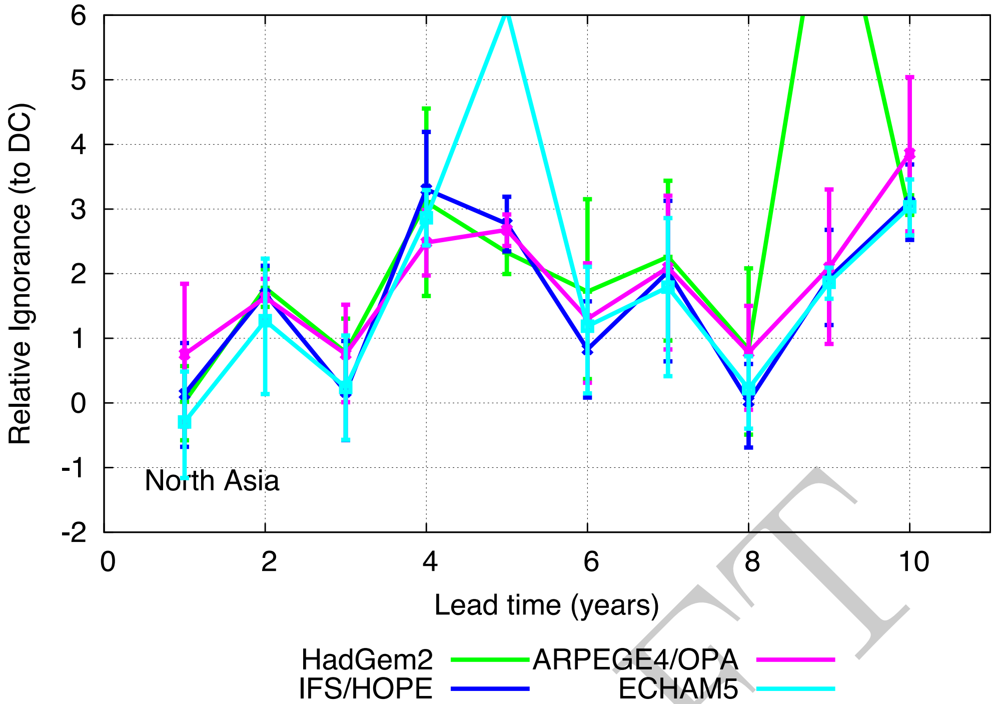


Fig. 16. Ignorance of the ENSEMBLES simulation models relative to the DC model for North Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

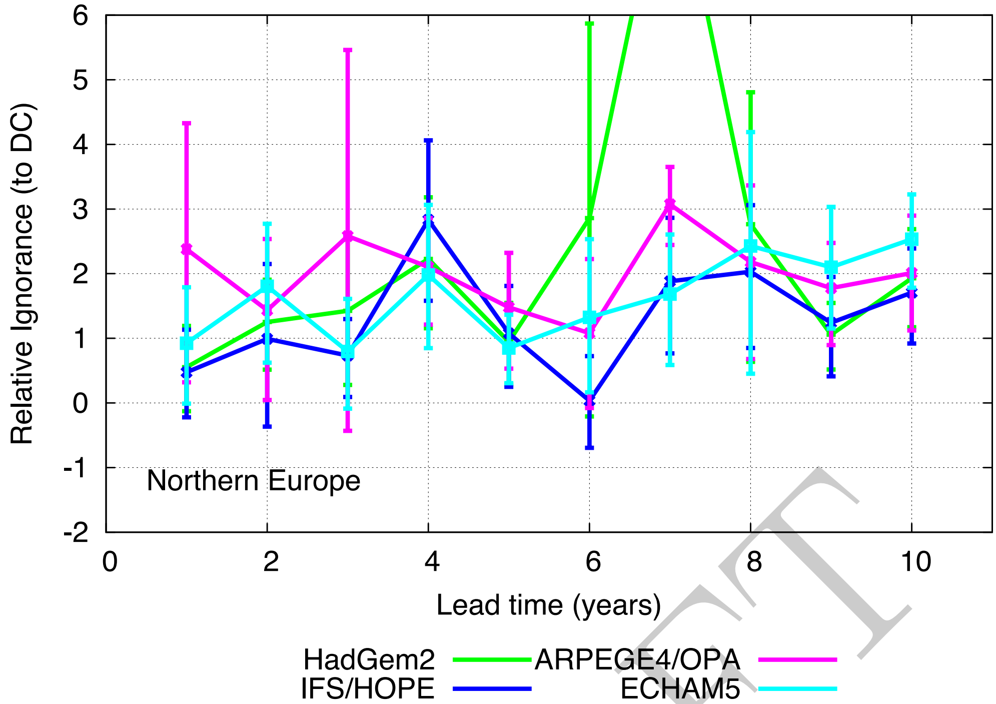


FIG. 17. Ignorance of the ENSEMBLES simulation models relative to the DC model for Northern Europe. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

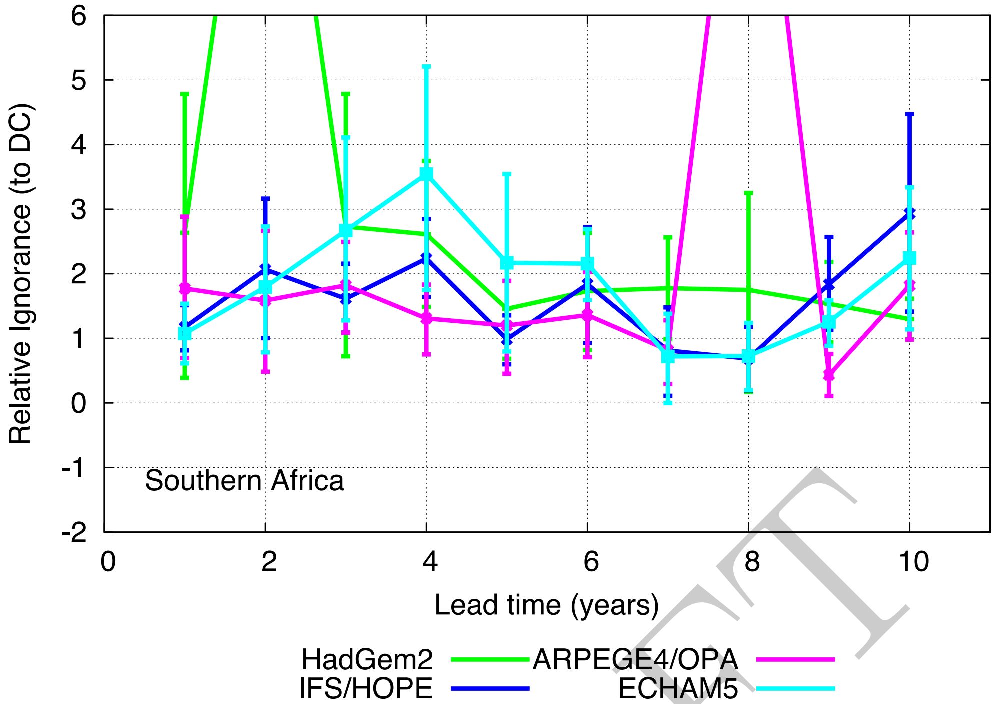


FIG. 18. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

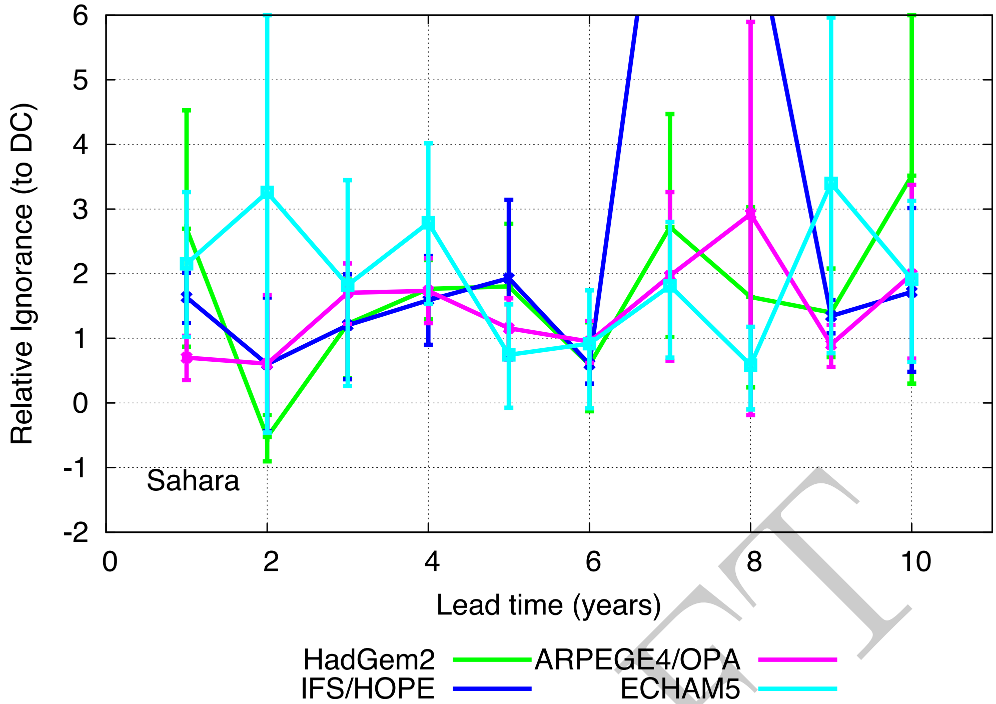


FIG. 19. Ignorance of the ENSEMBLES simulation models relative to the DC model for Sahara. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

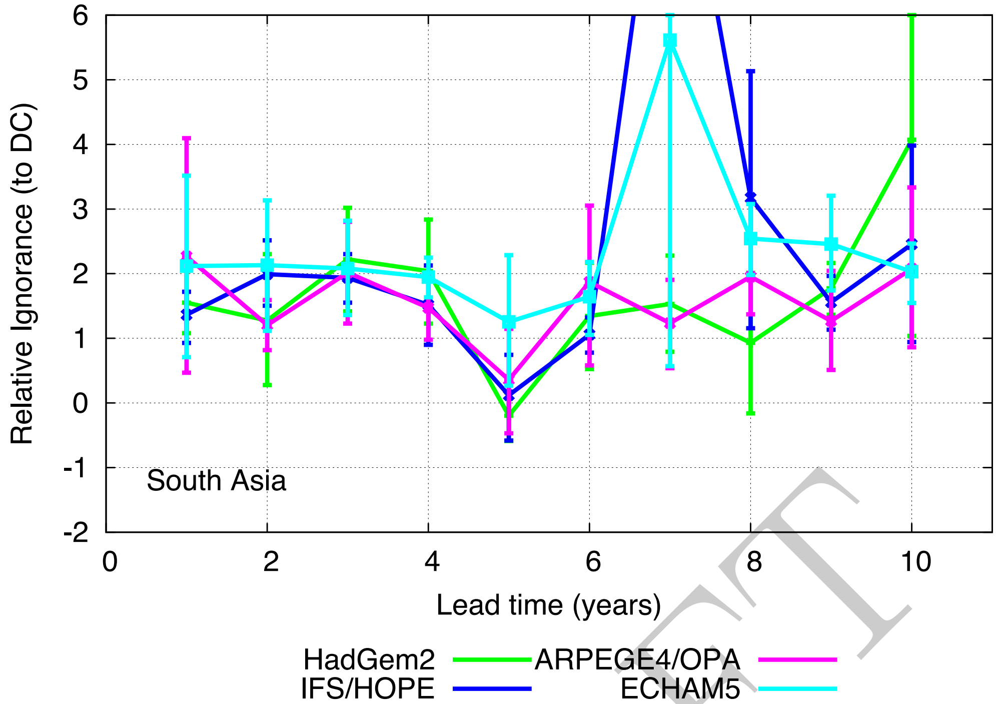


FIG. 20. Ignorance of the ENSEMBLES simulation models relative to the DC model for South Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

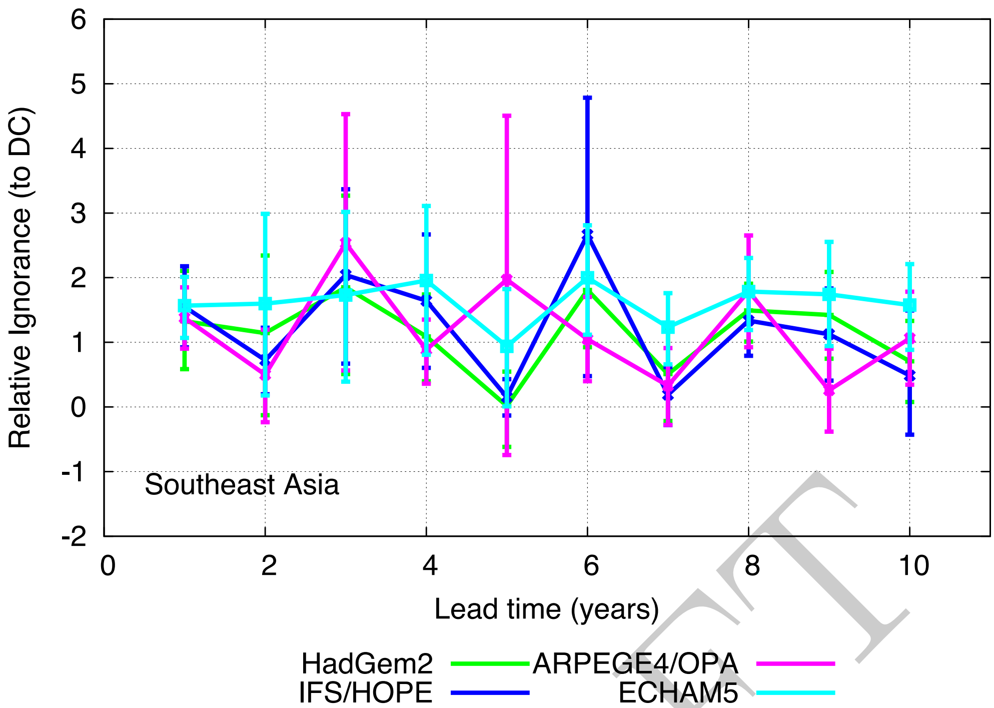


FIG. 21. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southeast Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

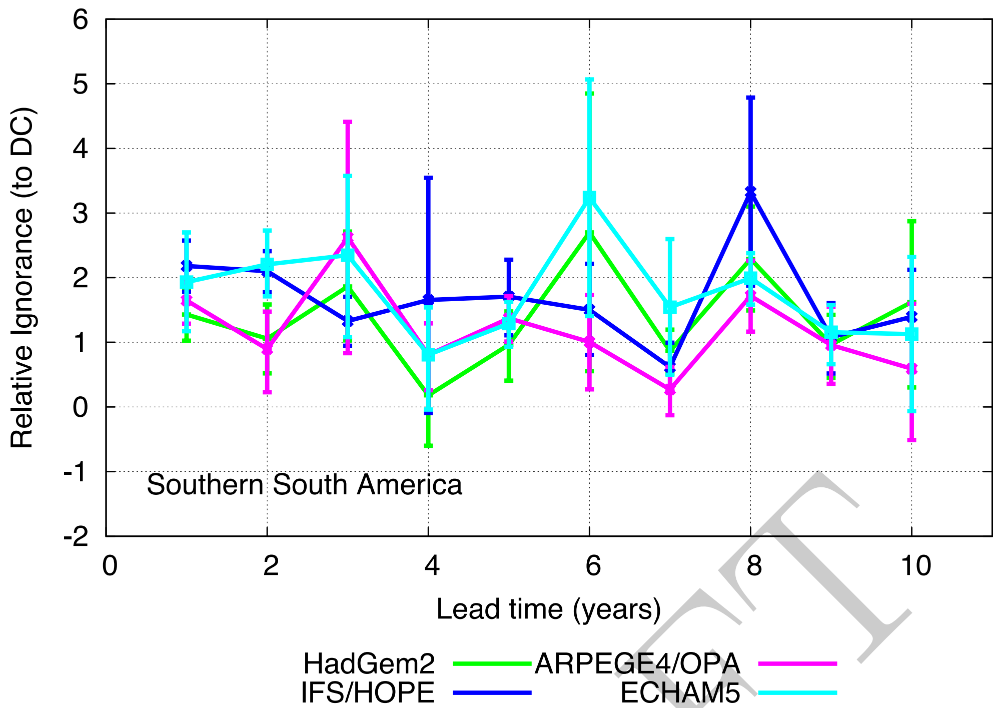


FIG. 22. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern South America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

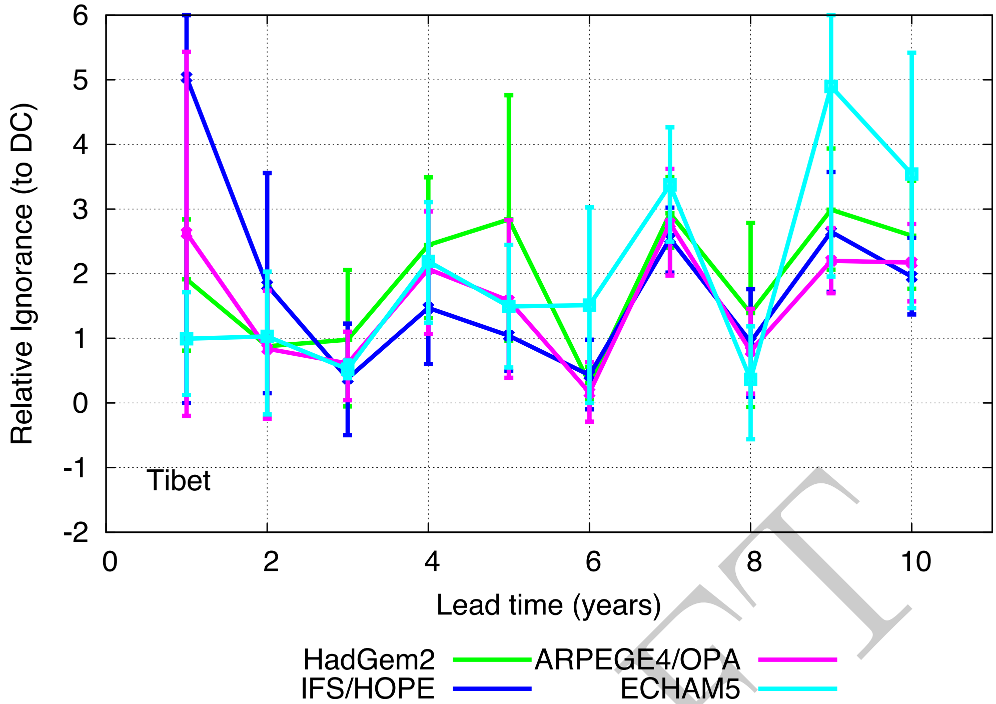


FIG. 23. Ignorance of the ENSEMBLES simulation models relative to the DC model for Tibet. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

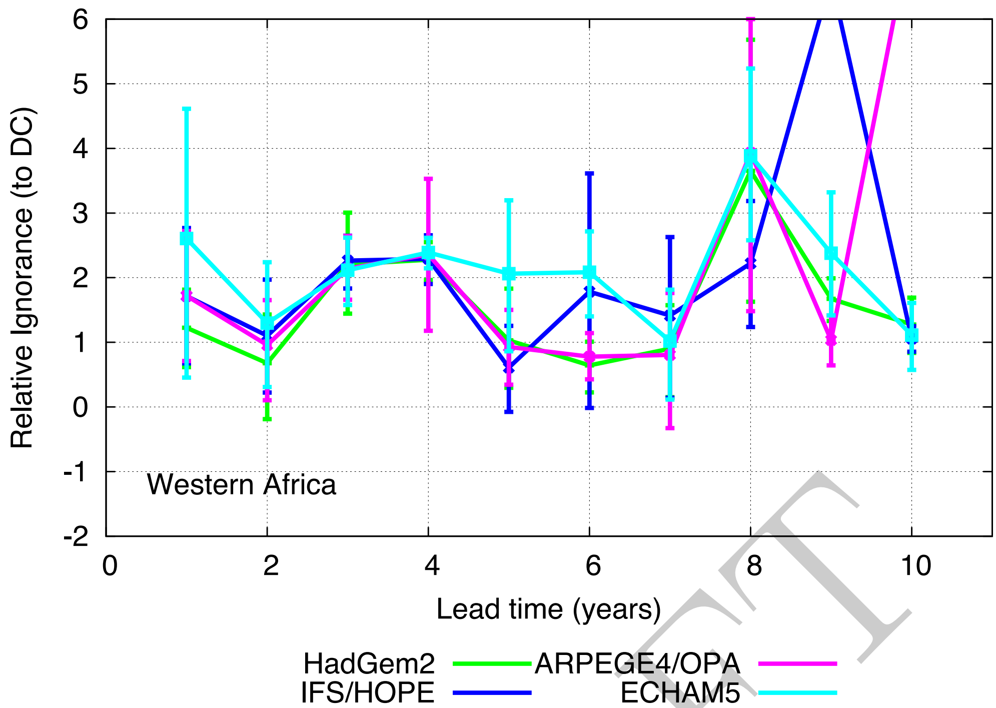


FIG. 24. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

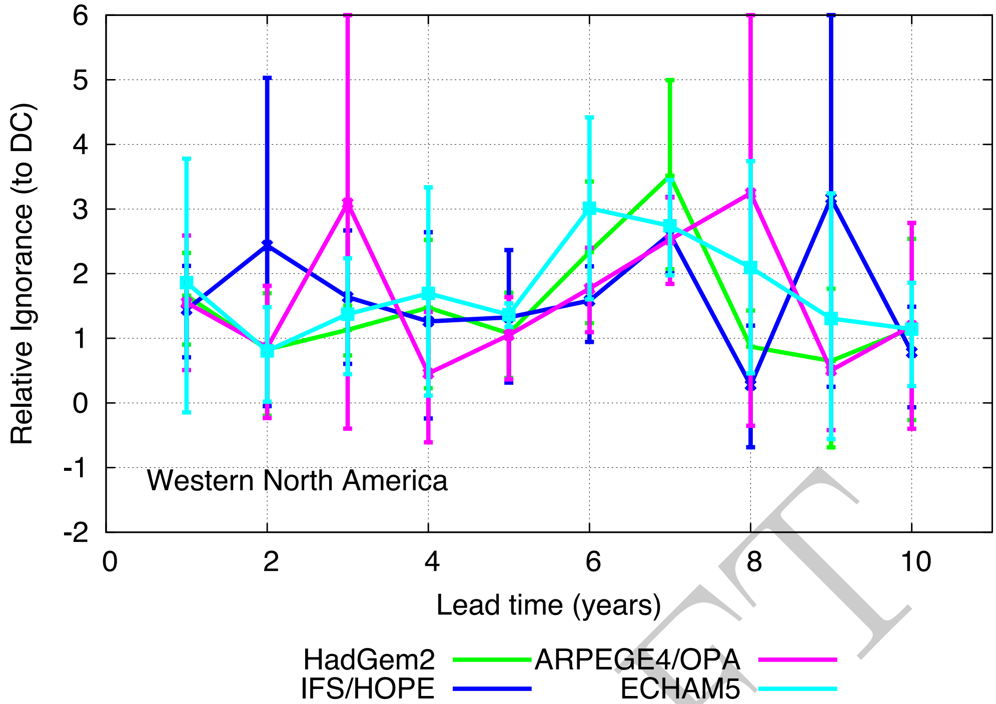


FIG. 25. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

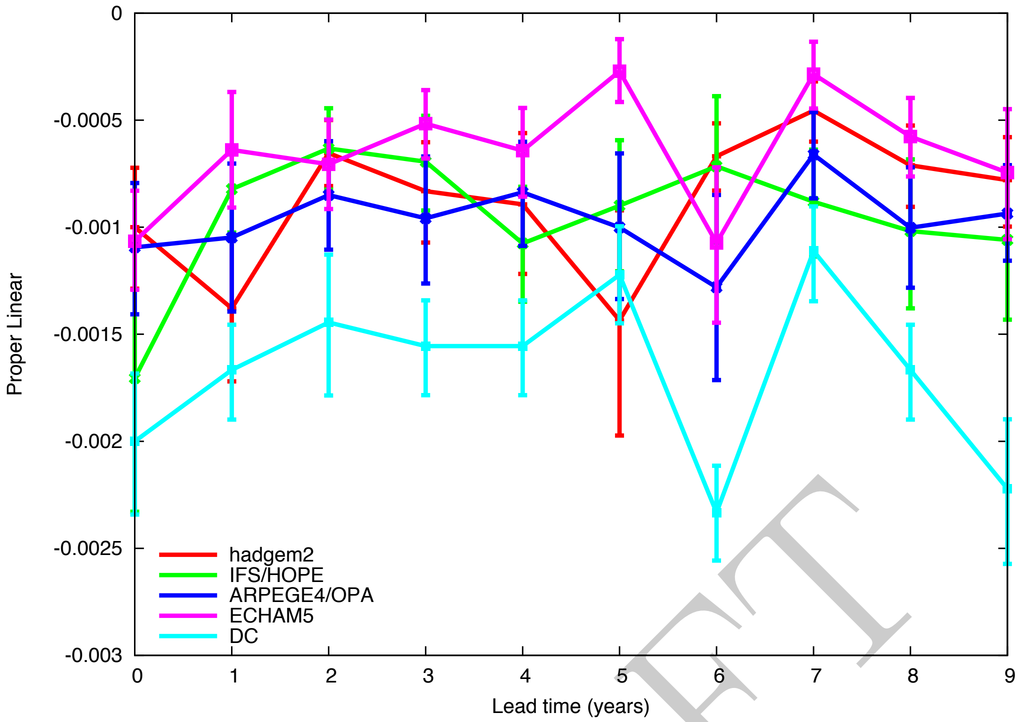


FIG. 26. Proper linear score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better foecasts. The DC model is shown to outperform the simulations models at most lead times.

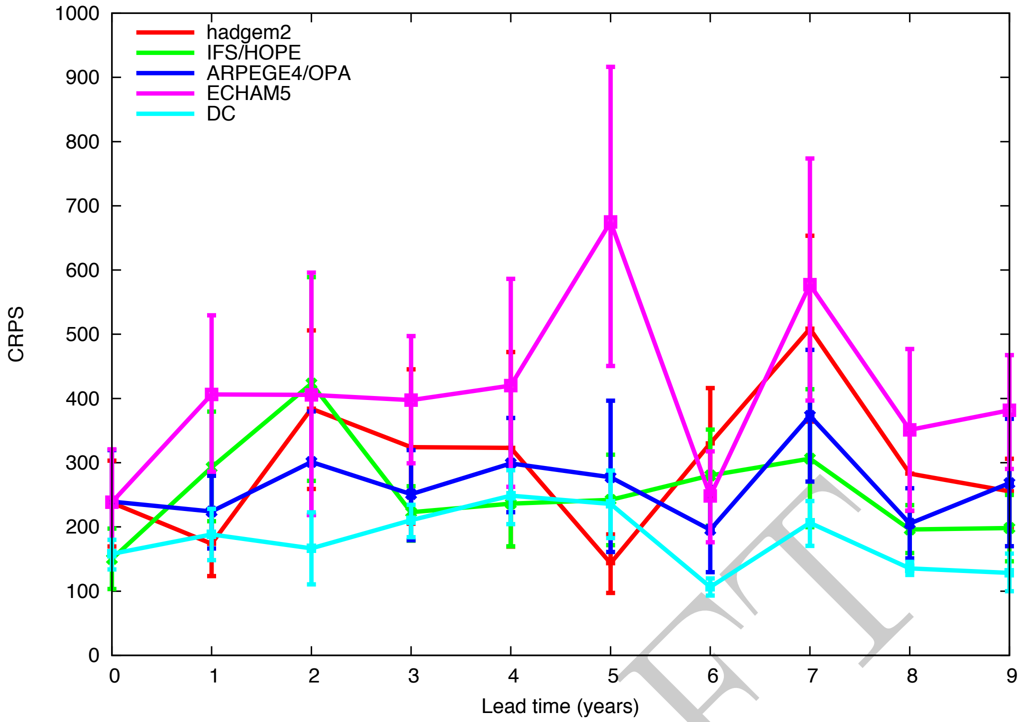


FIG. 27. CRPS score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better forecasts. The DC model is shown to outperform the simulations models at most lead times.

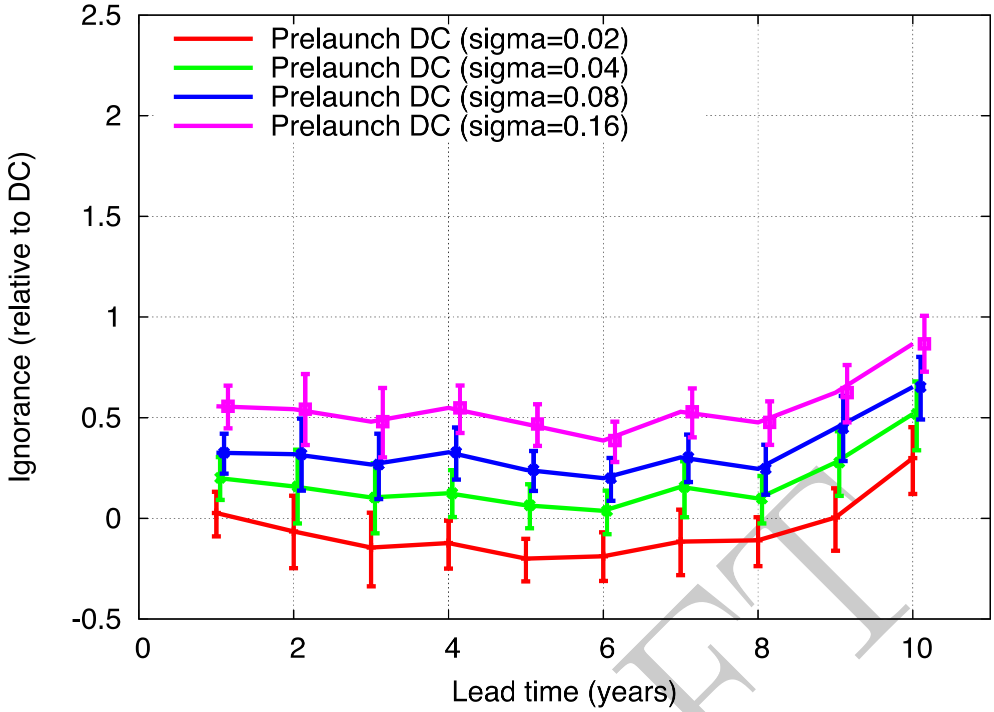


FIG. 28. Ignorance of the Prelaunch DC empirical model with kernel widths as labelled relative to the cross-validation DC model. Increasing the kernel width parameter from 0.02 to 0.16 results in a loss of skill of approximately half a bit, although for the kernel width value used in this paper (0.08)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

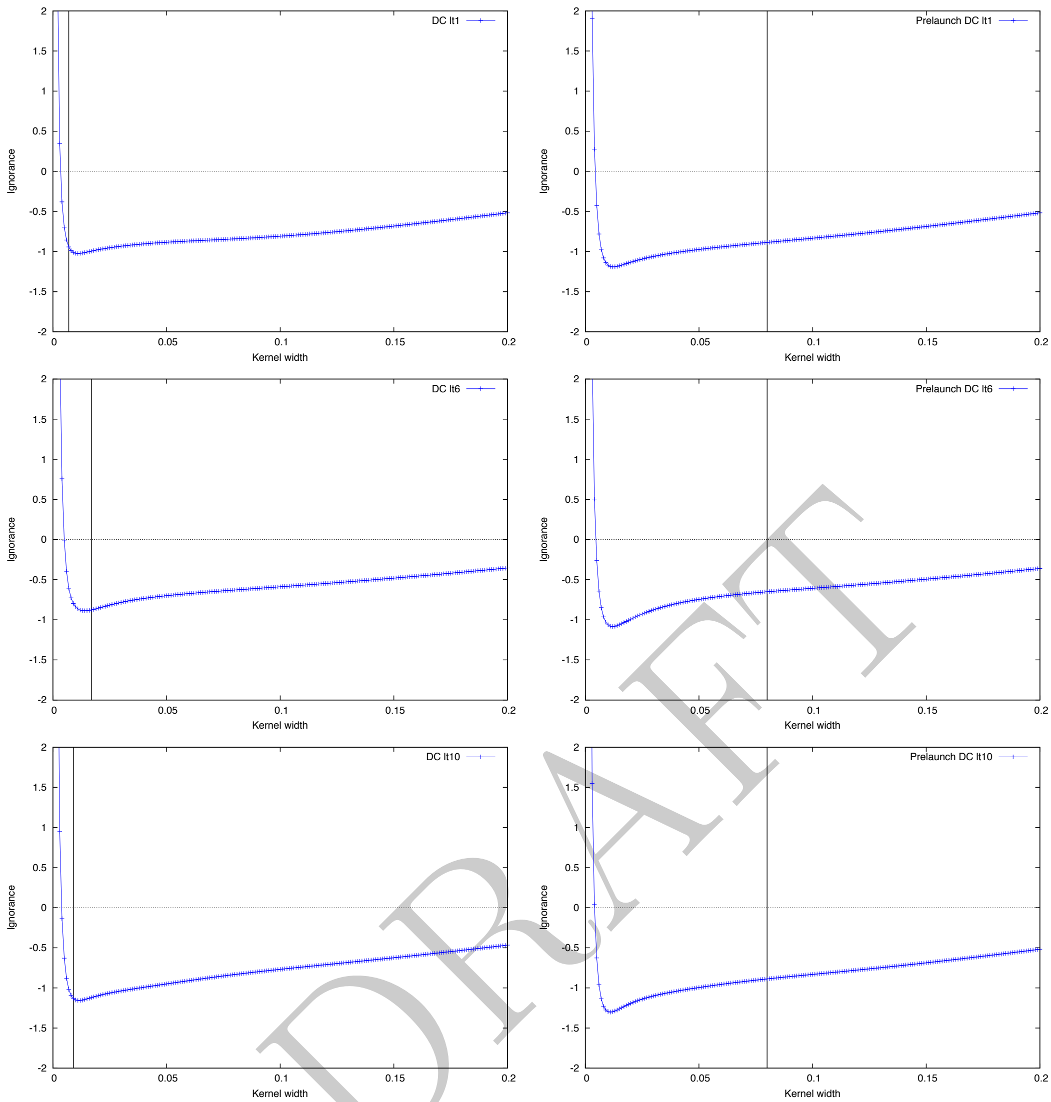




FIG. 29. Ignorance as a function of the kernel width parameter over the full set of hindcast simulations (*i.e.* with no cross-validation) for the DC (left panels) and Prelaunch DC (right panels) models at lead time one (a and b), six (c and d) and ten (e and f). The vertical bars in each case illustrate 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).

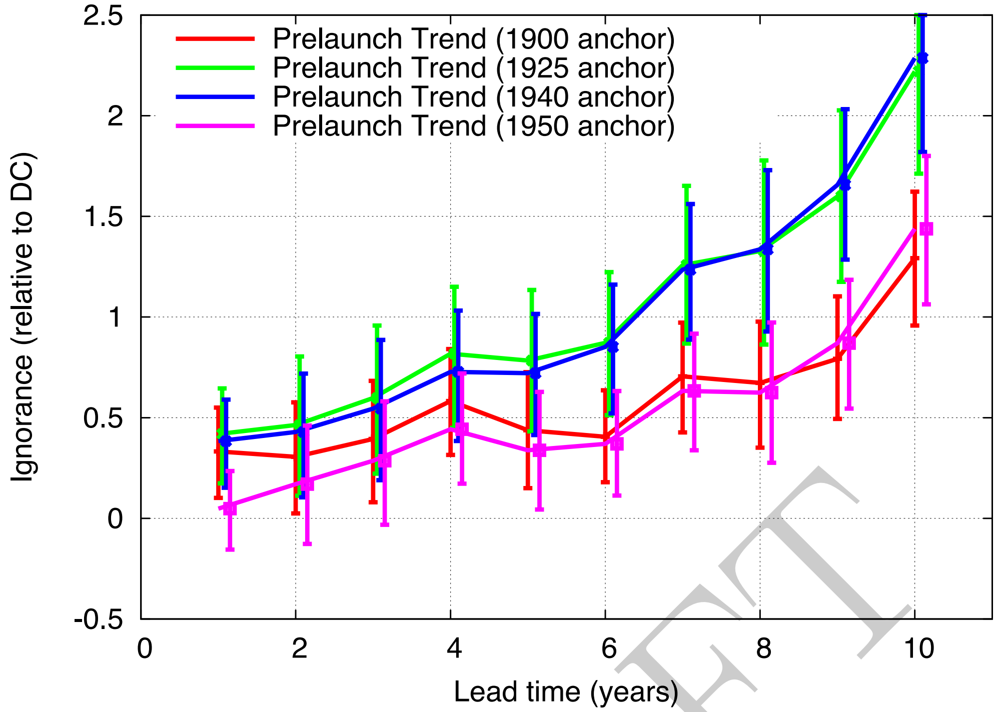


FIG. 30. Ignorance of the Prelaunch trend empirical model for different anchor times relative to the cross-validation DC model. Scores above zero indicate that DC outperforms the Prelaunch Trend model by up to half a bit at early lead times, and up to two bits (DC placing up to 4 times more probability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