



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 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>.

This working paper is intended to stimulate discussion within the research community and among users of research, and its content may have been submitted for publication in academic journals. It has been reviewed by at least one internal referee before publication. The views expressed in this paper represent those of the author(s) and do not necessarily represent those of the host institutions or funders.

1 **An evaluation of decadal probability forecasts**
2 **from state-of-the-art climate models**

3 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

4 21st October 2013

5 ABSTRACT

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

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

27 **1. Introduction**

28 State-of-the-art dynamical simulation models of the Earth's climate system¹ are often
29 used to make probabilistic predictions about the future climate and related phenomena with
30 the aim of providing useful information for decision support (Anderson et al. 1999; UK-
31 MetOffice 2011; Weigela and Bowlerb 2009; Alessandri et al. 2011; Hagedorn et al. 2005;
32 Hagedorn and Smith 2009; Meehl et al. 2009; Doblas-Reyes et al. 2010, 2011; IPCC 2007;
33 Reifen and Toumi 2009). Evaluating the performance of such predictions from a model,
34 or set of models, is crucial not only in terms of making scientific progress, but also in
35 determining how much information may be available to decision-makers via climate services.

¹Models that use physical principles to simulate the Earth's climate are often called general circula-
tion 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)

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

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

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

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

63 5 (CMIP5) decadal simulations (Taylor et al. 2009) is considered. The ENSEMBLES proba-
64 bility hindcasts are contrasted with forecasts from empirical models of the static climatology,
65 persistence and a “Dynamic Climatology” model developed for evaluating other dynamical
66 systems (Smith 1997; Binter 2011). Ensemble members are transformed into probabilistic
67 forecasts via kernel dressing (Bröcker and Smith 2008) and their quality quantified accord-
68 ing to several proper scoring rules (Bröcker and Smith 2006). The ENSEMBLES models do
69 not demonstrate significantly greater skill than that of an empirical Dynamic Climatology
70 model either for global mean temperature or for the land-based Giorgi region² temperatures
71 (Giorgi 2002).

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

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

²Giorgi 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.

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

88 **2. Decadal prediction systems**

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

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

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

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

111 whether initialising the model with observations at each forecast launch improves the skill of
112 decadal forecasts (Pohlmann et al. 2009; Hawkins et al. 2011; Smith et al. 2007; Keenlyide
113 et al. 2008; Smith et al. 2010; van Oldenborgh et al. 2012; Kim et al. 2012). At a more
114 basic level, the ability to provide useful decadal predictions using simulation models is yet
115 to be firmly established. Probabilistic hindcasts, based on simulations from Stream 2 of
116 the ENSEMBLES project (further details of which can be found in (Doblas-Reyes et al.
117 2010) and in the Appendix), do not demonstrate significantly more skill than that of simple
118 empirical models.

119 Figure 1 illustrates the 2-year running mean of simulated global mean temperature from
120 the four simulation models in the multi-model ensemble experiment of the ENSEMBLES
121 project over the full set of decadal hindcasts. Observations from the HadCRUT3 dataset
122 and ERA40 reanalysis are shown for comparison. HadCRUT3 is used as the outcome archive
123 for both the model evaluation and construction of the empirical model. Using ERA40 for
124 the verification instead of HadCRUT3 does not change the conclusions about the model skill
125 significantly (results not presented here). Global mean temperature is chosen for the analysis
126 as simulation models are expected to perform better over larger spatial scales (IPCC 2007).
127 Even at the global scale the raw simulation output is seen to differ from the observations both
128 in terms of absolute values, as well as in dynamics. Three of the four models display a sub-
129 stantial model drift away from the observed global mean temperature, the ECHAM5 model
130 is the exception. The fact that some of the models exhibit a substantial drift, but not oth-
131 ers, reflects the fact that different models employ different initialisation schemes (Keenlyide
132 et al. 2005). ECHAM5 both assimilates anomalies and forecasts anomalies. Assimilating
133 anomalies is intended to reduce model drift⁴ (Pierce et al. 2004); the remaining models are
134 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.

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

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

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

152 **3. Empirical models for decadal prediction**

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

160 and Smith 2009).

161 Empirical models based on historical observations cannot be expected to capture previ-
162 ously unobserved dynamics. Two empirical models typically used in forecast evaluation are
163 the climatological distribution and the persistence model. In the analysis below, a static
164 climatology defines a probabilistic distribution generated through the kernel dressing and
165 cross-validation procedures applied to the observational record (Bröcker and Smith 2008;
166 Hoeting et al. 1999), as outlined in section 4. Persistence forecasts are defined according
167 to a similar procedure, based on the last observation, persisted as a single ensemble mem-
168 ber for each launch. These models are not expected to prove ideal in a changing climate,
169 nevertheless information regarding the ability (or inability) of a simulation model to outper-
170 form these simple empirical models is of value. Alternative empirical models for probability
171 forecasts, more appropriate for a changing climate, define a Dynamic Climatology based on
172 ensemble random analogue prediction (eRAP) (Smith 1997; Paparella et al. 1997). Empir-
173 ical forecasts are also used as benchmarks for evaluating point forecasts of decadal climate
174 predictions (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012; Doblas-Reyes et al.
175 2010).

176 Analogue forecasting uses the current state (perhaps with other recent states (Smith
177 1997)) to define analogues within the observational record (van den Dool 1994; Lorenz 1963;
178 van den Dool 2007). A distribution based on images of each analogue state (the observation
179 immediately following the analogue state) then defines the ensemble forecast. Analogues may
180 be defined in a variety of ways, including near neighbours either in observation space or in a
181 delay reconstruction (Smith 1994, 1997). The ensemble members may be formed using the
182 complete set of available analogue states (Dynamic Climatology (Smith 1997; Binter 2011)),
183 or by selecting from the nearest neighbours at random, with the probability of selecting a
184 particular neighbour related to the distance (in the state space) between the prediction point
185 and the neighbour (the random analogue prediction method (Paparella et al. 1997)).

186 The dynamic climatologies constructed below provide l -step ahead forecast distributions

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

$$e_i = S_t + {}^l\Delta_i, \quad i = 1, \dots, N_l, \quad (1)$$

196
 197 where S_t is the initial condition at time t and ${}^l\Delta_i, i = 1, \dots, N_l$ is the set of l^{th} differences in
 198 the observational record. Figure 3 illustrates the DC model for global mean temperature,
 199 launched at five-year intervals, as in the ENSEMBLES hindcasts. A true out-of-sample
 200 forecast up to the year 2015, initialised to the observed global mean temperature in 2004,
 201 is also included. Each lead time 1 forecast is based on an ensemble of 48 members. In
 202 real-time forecasting $N_l = N - l$, while for cross-validation purposes the ensembles in figure
 203 3 use $N - l - 1$, omitting the Δ_j corresponding to the year being forecast. Thus at lead time
 204 9 each forecast is based on an ensemble of 40 members. The DC approach is shown below
 205 to outperform the ENSEMBLES models when forecasting global mean temperature.

206 4. Probability forecasts from ensembles

207 No forecast is complete without an estimate of forecast skill (Tennekes et al. 1987).
 208 Probability forecasts allow a complete description of the skill from an ensemble prediction
 209 system; they may be formed in several ways. The IPCC AR4 (IPCC 2007), for example,

210 defines a likely range subjectively, applying the ‘sixty-forty’ rule⁵ to the mean of the CMIP3
 211 model global mean temperatures in 2100. Insofar as the forecast-outcome archive is larger
 212 for decadal timescales, objective statistical approaches are more easily deployed.

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

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

217 where x^i is the i^{th} ensemble member, μ is the offset of the kernel mean (this offset may have
 218 a different value than the traditional “bias” term⁶) and σ is the kernel width. In this paper,
 219 the kernel, K , is taken to be a Gaussian function,

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

220 The kernel parameters are fitted by minimising a chosen skill score (Jolliffe and Stephenson
 221 2003) while avoiding information contamination⁷.

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

$$S(p(y), Y) = -\log_2(p(Y)), \quad (4)$$

⁵In 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.

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

⁷Information 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.

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

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

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

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

230 Appropriate reference forecasts will depend on the task at hand: they may include a
 231 static climatological distribution, a dynamic climatology, another simulation model or em-
 232 pirical model. The relative Ignorance quantifies (in terms of bits) the additional information
 233 provided by forecasts from one model, above that of the reference. A relative Ignorance
 234 score of $S_{rel}(p_1(y), p_2(y), Y) = -1$ means that the model forecast places, on average, twice
 235 (that is 2^1) the probability mass on the verification than the reference forecast. Similarly, a
 236 score of

237 $S_{rel}(p_1(y), p_2(y), Y) = -1/2$ means $\sim 41\%$ (that is $2^{1/2}$) more probability mass on average.

238 In section 5, the static climatology, a persistence forecast and the DC model are chosen
 239 as references to measure performance against the ENSEMBLES simulation models. The
 240 parameters used to construct each empirical model forecast are each estimated under true-
 241 cross-validation: the forecast target decade is omitted from consideration.

242 The ENSEMBLES forecast-outcome archive contains at most nine forecast-outcome
 243 pairs. That is, there are only nine forecast launch dates, each with a maximum lead time of
 244 ten years. Outside of true out-of-sample evaluation it is difficult not to overfit the forecast
 245 and dressing parameters used to generate probability forecasts; the details of cross-validation

246 can have a large impact. Extending the typical leave-some-out fitting protocol (Hastie et al.
247 2001; Bröcker and Smith 2008) to include the kernel dressing procedure reduces the sample
248 size of the forecast-outcome archive from eight to seven pairs. This ‘true leave-some-out’
249 procedure (Smith et al. 2013) will necessarily increase the sampling uncertainty, reflected
250 through bootstrap resampling. In the case of the ENSEMBLES forecasts, adopting a true
251 leave-some-out procedure reduces the apparent significance of the results; failing to introduce
252 such a procedure, however, risks both information contamination and the suggestion that
253 there is more skill than is to be expected in the simulation models. The most appropriate
254 path cannot be determined with confidence until additional data becomes available.

255 5. Results

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

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

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

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

298 relative skill axis, clouding the significance of the result.

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

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

318 The poor performance of the ECHAM5 simulation model might at first appear as a
319 surprise, since the ensemble members from this model appear to be relatively close to the
320 target values in figure 1. Note, however, that ECHAM5 initialises (and thus forecasts)
321 model anomalies, not physical temperatures; the model forecasts then yield forecast model
322 anomalies. In this case then, the systematic error of the model is partially accounted for
323 when the model forecast anomalies are translated back into physical temperatures. The offset
324 applied within the kernel dressing procedure levels the playing field by accounting for the

325 systematic errors in the other simulation models; the figures indicate that while ECHAM5
326 may suffer less model drift due to this process (Keenlyside et al. 2005) it does not produce
327 more skillful probability forecasts than the other ENSEMBLES simulation models.

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

336 While hindcast experiments can never provide true “out of sample” evaluation of a fore-
337 casting system, it is possible to deny empirical models access to data observed after each
338 launch date. In addition to the denial of what were effectively future observations, it is also
339 necessary to illustrate that the skill of these *Prelaunch* empirical models⁸ does not depend
340 sensitively on parameter tuning, as it is implausible that such tuning could have been done
341 in real-time. The results reported below are robust to variations in the free parameters in
342 the Prelaunch DC model (see Supplementary Material).

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

⁸Arguably 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. “Prelaunch” 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.

348 to decreasing this width by more than an order of magnitude. A Prelaunch Trend model was
349 also constructed to determine if the observed skill was due to a linear trend. The Prelaunch
350 Trend model simply extends the linear fit to the observations from a fixed start-date (say,
351 1950) to the launch date, and then uses the standard deviation of the residuals as the kernel
352 width. The Prelaunch DC is more skillful than the Prelaunch Trend model, as shown in figure
353 10. This result is robust to changing the start-date back towards 1900 (see Supplementary
354 Material). It is important to stress that this trend model is not being advocated as a
355 candidate empirical model, but only to address the specific question of whether the skill of
356 the DC model comes only from the observed trend in global mean temperature. Much more
357 effective methods for estimating statistical time-series models are available in this context
358 (see for example (Fildes and Kourentzes 2011)).

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

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

381 **6. Conclusions**

382 The quality of decadal probability forecasts from the ENSEMBLES simulation models
383 has been compared with that of reference forecasts from several empirical models. In general,
384 the Stream 2 ENSEMBLES simulation models demonstrate less skill than the empirical DC
385 model across the range of lead times from one to ten years. The result holds for a variety of
386 proper scoring rules including Ignorance (Good 1952), the Proper Linear Score (PL) (Jolliffe
387 and Stephenson 2003) and the continuous ranked probability score (CRPS) (Bröcker and
388 Smith 2006). A similar result holds on smaller spatial scales for the Giorgi Regions (see
389 Supplementary Material). These new results for probability forecasts are consistent with
390 evaluations of root-mean-square errors of decadal simulation models with other reference
391 point forecasts (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012; Weisheimer et al.
392 2009). The DC probability forecasts often place up to 4 bits more information (or 2^4 times
393 more probability mass) on the observed outcome than the ENSEMBLES simulation models.

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

400 empirical models) on longer time scales.

401 The evaluation and comparison of decadal forecasts will always be hindered by the rela-
402 tively small samples involved when contrasted with the case of weather forecasts; the decadal
403 forecast-outcome archive currently considered is only half a century in duration. Advances
404 both in modelling and in observation, as well as changes in the Earth’s climate, are likely to
405 mean the relevant forecast-outcome archive will remain small. One improvement that could
406 be made to clarify the skill of the simulation models is to improve the experimental design
407 of hindcasts, in particular to increase the ensemble size used. For the ENSEMBLES models,
408 each simulation ensemble consisted of only three members launched at five year intervals.
409 Larger ensembles and more frequent forecast launch dates can ease the evaluation of skill
410 without waiting for the forecast-outcome archive to grow larger⁹.

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

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

⁹As 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.

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

431 **Acknowledgments**

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

442 **Appendix: The Stream 2 ENSEMBLES decadal hind-** 443 **cast experiments**

444 The set of decadal hindcast experiments from Stream 2 of the ENSEMBLES project
445 simulations (Doblas-Reyes et al. 2010) have a similar experimental design to the seasonal

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

459 460 **REFERENCES**

461 Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Deque, P. Rogel, and A. Weisheimer,
462 2011: Evaluation of probabilistic quality and value of the ensembles multimodel seasonal
463 forecasts: Comparison with demeter. *Monthly Weather Review*, **139**, 581–607.

464 Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999:
465 Present-day capabilities of numerical and statistical models for atmospheric extratropi-
466 cal seasonal simulation and prediction. *Bulletin of the American Meteorological Society*,
467 **80** (7).

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

- 470 Binter, R., 2011: Applied probabilistic forecasting. Ph.D. thesis, London School of Economics
471 and Political Science.
- 472 Bröcker, J. and L. A. Smith, 2006: Scoring probabilistic forecasts: The importance of being
473 proper. *Weather and Forecasting*, **22**, 382–388.
- 474 Bröcker, J. and L. A. Smith, 2008: From ensemble forecasts to predictive distributions.
475 *Tellus A*, **60** (4), 663–678.
- 476 Colman, A. and M. Davey, 2003: Statistical prediction of global sea-surface temperature
477 anomalies. *International Journal of Climatology*, **23** (956), 1677–1697.
- 478 Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer, 2011: Decadal
479 climate prediction with the european centre for medium-range weather forecasts coupled
480 forecast system: Impact of ocean observations. *Journal of Geophysical Research - Atmo-*
481 *spheres*, **116** (D19111).
- 482 Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Fore-
483 cast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. *Technical*
484 *Memorandum ECMWF*, **621**.
- 485 Fildes, R. and N. Kourentzes, 2011: Validation and forecasting accuracy in models of climate
486 change. *International Journal of Forecasting*, **27** (4).
- 487 Giorgi, F., 2002: Variability and trends of sub-continental scale surface climate in the twen-
488 tieth century. part i: observations. *Climate Dynamics*, **18**, 675–691.
- 489 Good, I. J., 1952: Rational decisions. *Journal of the Royal Statistical Society*, **XIV** (1),
490 107–114.
- 491 Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success
492 of multi-model ensembles in seasonal forecasting. part i: Basic concept. *Tellus*, **A57**, 219–
493 233.

- 494 Hagedorn, R. and L. A. Smith, 2009: Communicating the value of probabilistic forecasts
495 with weather roulette. *Meteorological Applications*, **16 (2)**, 143–155.
- 496 Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The elements of statistical learning*.
497 Springer, New York.
- 498 Hawkins, E., J. Robson, R. Sutton, D. Smith, and N. Keenlyside, 2011: Evaluating the
499 potential for statistical decadal predictions of sea surface temperature with a perfect model
500 approach. *Climate Dynamics*, 1–15.
- 501 Hewitt, C. D. and D. J. Griggs, 2004: Ensembles-based predictions of climate and their
502 impacts. *Eos, Transactions American Geophysical Union*, **85 (52)**, 566.
- 503 Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model
504 averaging: A tutorial. *Statistical Science*, **14 (4)**.
- 505 IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working*
506 *Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate*
507 *Change [S. Solomon and D. Qin and M. Manning and Z. Chen and M. Marquis and K.*
508 *B. Averyt and M. Tignor and H. L. Miller (eds.)].* 996 pp, Cambridge University Press,
509 Cambridge, United Kingdom and New York, NY USA.
- 510 Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast verification: A practitioner's guide in*
511 *atmospheric science*. John Wiley and Sons Ltd.
- 512 Keenlyside, N. S., M. Latif, M. Botzet, J. Jungclaus, and U. Schulzweida, 2005: A coupled
513 method for initializing el nino southern oscillation forecasts using sea surface temperature.
514 *Tellus*, **57A**, 340–356.
- 515 Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing
516 decadal-scale climate prediction in the north atlantic sector. *Nature*, **453 (06921)**, 84–88.

- 517 Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate
518 change prediction in multi-model cmip5 decadal hindcasts. *Geophysical Research Letters*,
519 **39 (L10701)**.
- 520 Krueger, O. and J.-S. von Storch, 2011: A simple empirical model for decadal prediction.
521 *Journal of Climate*, **24**, 1276–1283.
- 522 Laepple, T., S. Jewson, and K. Coughlin, 2008: Interannual temperature predictions using
523 the cmip3 multi-model ensemble mean. *Geophysical Research Letters*, **35 (L10701)**.
- 524 Lee, T. C. K., F. W. Zwiers, X. Zhang, and M. Tsao, 2006: Evidence of decadal climate pre-
525 diction skill resulting from changes in anthropogenic forcing. *Journal of Climate*, **19 (20)**,
526 5305–5318.
- 527 Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20 (2)**,
528 130–141.
- 529 Meehl, G. A., et al., 2009: Decadal prediction: Can it be skillful? *Bulletin of the American*
530 *Meteorological Society*, **90**, 1467–1485.
- 531 Paparella, F., A. Provenzale, L. A. Smith, C. Taricco, and R. Vio, 1997: Local random
532 analogue prediction of nonlinear processes. *Physics Letters A*, **235 (3)**, 233–240.
- 533 Pierce, D. W., T. P. Barnett, R. Tokmakian, A. Semtner, M. Maltrud, J. A. Lysne, and
534 A. Craig, 2004: The acpi project, element 1: Initialising a coupled climate model from
535 observed conditions. *Climatic Change*, **62 (1)**, 13–28.
- 536 Pohlmann, H., J. H. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate
537 predictions with gecco oceanic synthesis: Effects on the north atlantic. *Journal of Climate*,
538 **22**, 3926–3938.
- 539 Reifen, C. and R. Toumi, 2009: Climate projections: Past performance no guarantee of
540 future skill? *Geophysical Research Letters*, **36 (L13704)**.

- 541 Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy,
542 2007: Improved surface temperature prediction for the coming decade from a global climate
543 model. *Science*, **317**, 796–799.
- 544 Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and
545 A. A. Scaife, 2010: Skilful multi-year predictions of atlantic hurricane frequency. *Nature*
546 *Geoscience*, **3** (1004), 846–849.
- 547 Smith, L. A., 1992: Identification and prediction of low-dimensional dynamics. *Physica D*,
548 **58** (1-4), 50–76.
- 549 Smith, L. A., 1994: Local optimal prediction: Exploiting strangeness and the variation of sen-
550 sitivity to initial condition. *Philosophical Transactions of the Royal Society A*, **348** (1688),
551 371–381.
- 552 Smith, L. A., 1997: The maintenance of uncertainty. *Proc. International School of Physics*
553 *“Enrico Fermi”*, **Course CXXXIII**, 177–246, Societ’a Italiana di Fisica, Italy.
- 554 Smith, L. A., 2002: What might we learn from climate forecasts? *Proceedings of the National*
555 *Academy of Science*, **4** (99), 2487–2492.
- 556 Smith, L. A., H. Du, F. Niehörster, and E. B. Suckling, 2013: An evaluation of probabilistic
557 skill from ensemble seasonal forecasts. *submitted to the Quarterly Journal of the Royal*
558 *Meteorological Society*.
- 559 Stockdale, T. N., 1997: Coupled ocean-atmosphere forecasts in the presence of climate drift.
560 *Monthly Weather Review*, **125**, 809–818.
- 561 Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the cmip5 experimental
562 design. http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.
- 563 Tennekes, H., A. P. M. Baede, and J. D. Opsteegh, 1987: Forecasting forecast skill. *Proceed-*
564 *ings ECMWF Workshop on Predictability, ECMWF, Reading, UK*, 277–302.

- 565 Troccoli, A. and T. N. Palmer, 2007: Ensemble decadal predictions from analysed initial
566 conditions. *Philosophical Transactions of the Royal Society A*, **365**, 2179–2191.
- 567 UK-MetOffice, 2011: 3-month outlook for uk contingency planning. [http://www.
568 metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_user_guidance-150.pdf](http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_user_guidance-150.pdf).
- 569 Unger, D., H. van den Dool, E. O’Lenic, and D. Collins, 2009: Ensemble regression. *Monthly
570 Weather Review*, **137 (2365-2379)**.
- 571 van den Dool, H. M., 1994: Long-range weather forecasts through numerical and empirical
572 methods. *Dynamics of Atmospheres and Oceans*, **20 (3)**, 247–270.
- 573 van den Dool, H. M., 2007: *Empirical methods in short-term climate prediction*. Oxford
574 University Press.
- 575 van Oldenborgh, G. J., M. Balmaseda, L. Ferranti, T. Stockdale, and D. Anderson, 2005:
576 Evaluation of atmospheric fields from the ecnwf seasonal forecasts over a 15-year period.
577 *Journal of Climate*, **18**, 3250–3269.
- 578 van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal
579 prediction skill in a multi-model ensemble. *Climate Dynamics*, **38 (7-8)**, 1263–1280.
- 580 Weigela, A. P. and N. E. Bowlerb, 2009: Can multi-model combination really enhance predic-
581 tion skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological
582 Society*, **135**, 535–539.
- 583 Weisheimer, A., et al., 2009: Ensembles - a new multi-model ensemble for seasonal-to-
584 annual predictions: Skill and progress beyond demeter in forecasting tropical pacific ssts.
585 *Geophysical Research Letters*, **36 (L21711)**.
- 586 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences, 3rd Edition*, Vol. 100.
587 Academic Press.

List of Figures

- 588
- 589 1 Global mean temperature (2 year running mean) for the four forecast systems
590 - HadGem2 (UKMO), IFS/HOPE (ECMWF), ARPEGE4/OPA (CERFACS)
591 and ECHAM5 (IFM-GEOMAR) - that form Stream 2 of the ENSEMBLES
592 decadal hindcast simulations (Doblas-Reyes et al. 2010). HadCRUT3 obser-
593 vations and ERA40 reanalysis are also shown for comparison. Note that the
594 scale on the vertical axis for the ARPEGE4/OPA model is different to the
595 other three panels, reflecting the larger bias in this model. 28
- 596 2 Mean forecast error as a function of lead time across the set of decadal hind-
597 casts for each of the ENSEMBLES simulation models as labelled. Note that
598 the scale on the vertical axis for the ARPEGE4/OPA model is different to
599 the other three panels, reflecting the larger bias in this model. 29
- 600 3 Dynamic climatology (DC) over the period of the ENSEMBLES hindcasts
601 (figure 1). HadCRUT3 (from which the DC model is constructed) is shown
602 for comparison. 30
- 603 4 Forecast distributions for HadGem2 (UKMO) for the 5-95th percentile. The
604 HadCRUT3 observed temperatures are shown in blue. The forecasts are ten
605 years long and launched every five years, and so the fan charts would overlap;
606 to avoid this they are presented on two panels. The top (bottom) panel
607 illustrates forecasts launched in ten year intervals from 1960 (1965). 31
- 608 5 Forecast distribution for every fifth launch from the Dynamic Climatology
609 (DC) model for the 5-95th percentile. The HadCRUT3 observed temperatures
610 are shown in blue. The forecasts are ten years long and launched every five
611 years, and so the fan charts would overlap; to avoid this they are presented
612 on two panels. The top (bottom) panel illustrates forecasts launched in ten
613 year intervals from 1960 (1965). 32

- 614 6 Ignorance as a function of lead time for each of the four ENSEMBLES hindcast
615 simulation models and the DC model relative to the static climatology. The
616 bootstrap resampling intervals are illustrated at the 10-90th percent level. The
617 DC model is shown to be significantly more skillful than static climatology
618 at all lead times, whereas the ARPEGE4/OPA and IFS/HOPE models are
619 significantly more skillful than static climatology at early lead times. 33
- 620 7 Probability density for the static climatology used in the paper with obser-
621 vations over the period 1960-2010 (from HadCRUT3) illustrated as points on
622 the x-axis for reference. 34
- 623 8 Ignorance of the ENSEMBLES models and DC relative to persistence fore-
624 casts as a function of lead time. The DC model has negative relative Ignorance
625 scores up to 6 years ahead, indicating it is significantly more skillful than per-
626 sistence forecasts at early lead times. The ENSEMBLES models tend to have
627 positive scores, particularly at longer lead times, with bootstrap resampling
628 intervals that overlap with the zero skill line. The bootstrap resampling in-
629 tervals are illustrated at the 10-90th percent level. 35
- 630 9 Ignorance of the ENSEMBLES models relative to DC as a function of lead
631 time. The bootstrap resampling intervals are illustrated at the 10-90th percent
632 level. Note that the simulation models tend to have positive scores (less skill)
633 than the DC model at every lead time. 36
- 634 10 Ignorance of the Prelaunch DC and Prelaunch Trend models relative to the
635 standard DC model as a function of lead time. The HadGem2 model from
636 ENSEMBLES is also shown. It is shown that the Prelaunch DC model is
637 not significantly less skillful than the standard DC model and is robust to
638 variations in parameter tuning. The Prelaunch linear trend model is, however,
639 generally shown to be less skillful than the standard DC model. The bootstrap
640 resampling intervals are illustrated at the 10-90th percent level. 37

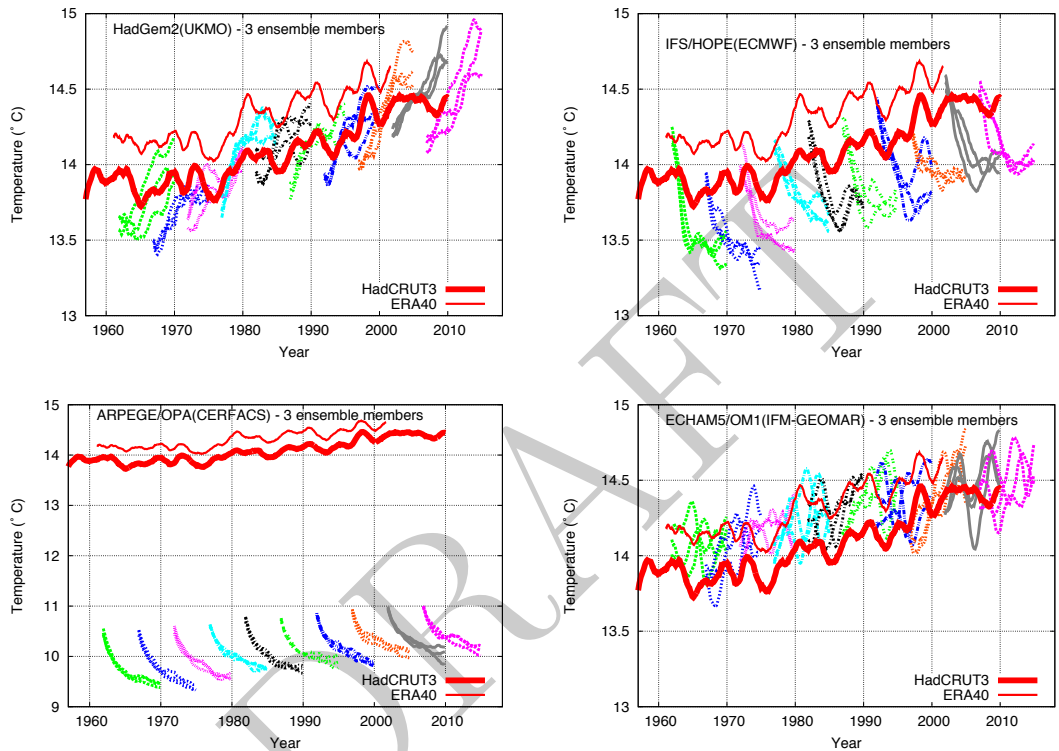


FIG. 1. Global mean temperature (2 year running mean) for the four forecast systems - HadGem2 (UKMO), IFS/HOPE (ECMWF), ARPEGE4/OPA (CERFACS) and ECHAM5 (IFM-GEOMAR) - that form Stream 2 of the ENSEMBLES decadal hindcast simulations (Doblas-Reyes et al. 2010). HadCRUT3 observations and ERA40 reanalysis 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.

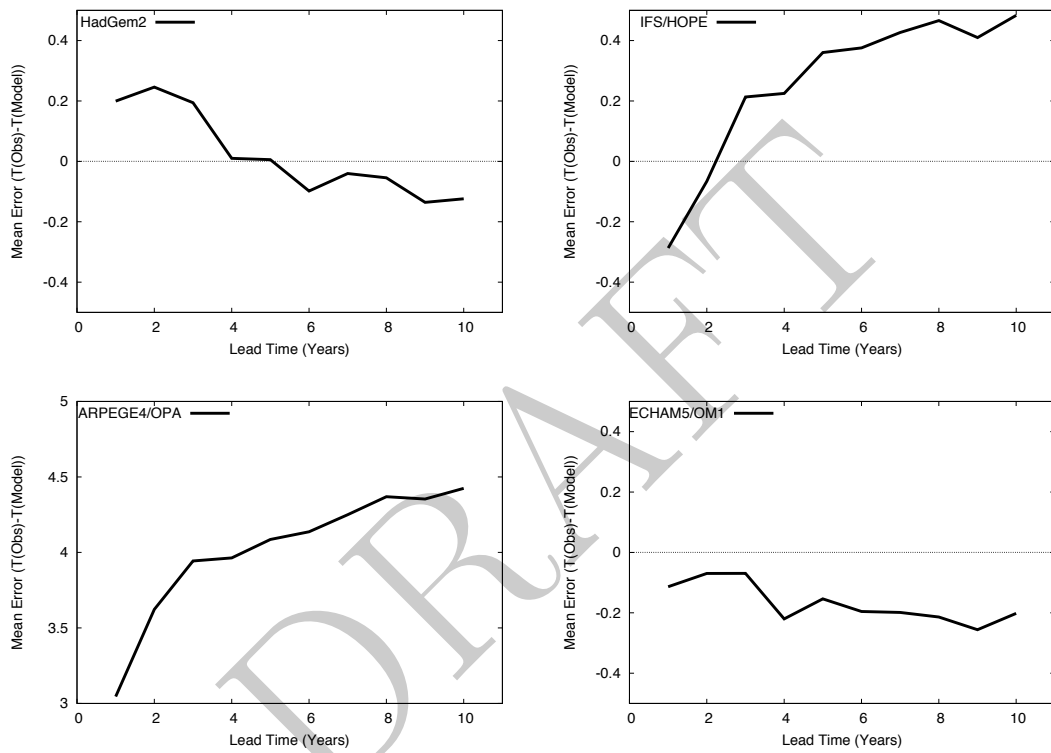


FIG. 2. Mean forecast error as a function of lead time across the set of decadal hindcasts for each of the ENSEMBLES simulation models as labelled. Note that the scale on the vertical axis for the ARPEGE4/OPA model is different to the other three panels, reflecting the larger bias in this model.

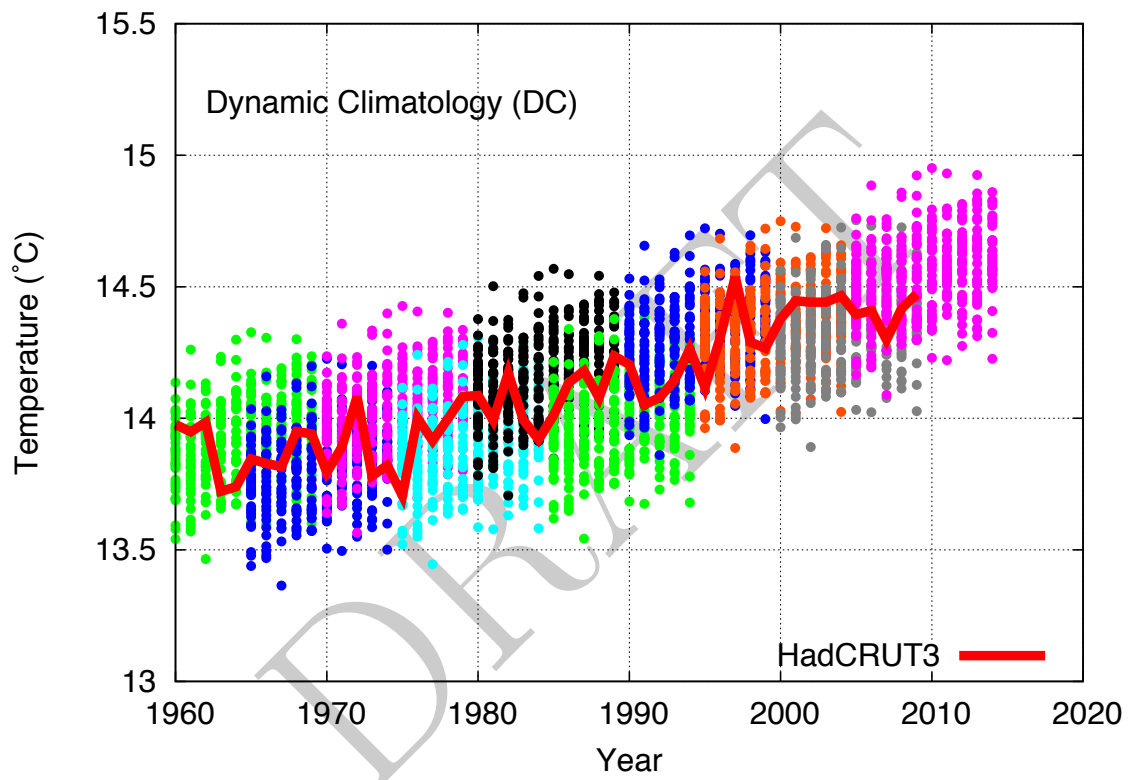


FIG. 3. Dynamic climatology (DC) over the period of the ENSEMBLES hindcasts (figure 1). HadCRUT3 (from which the DC model is constructed) is shown for comparison.

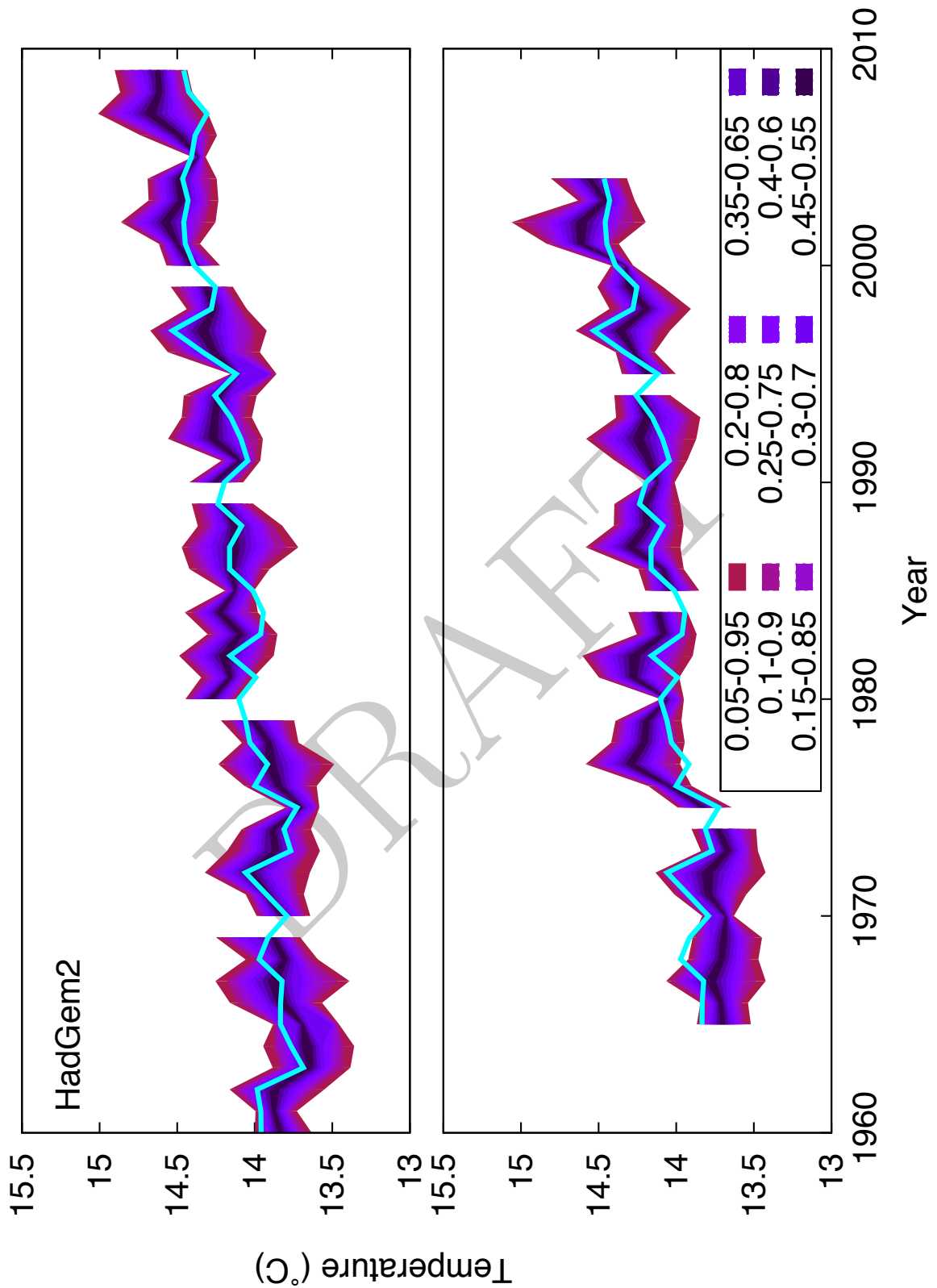


FIG. 4. Forecast distributions for HadGem2 (UKMO) for the 5-95th percentile. The Had-CRUT3 observed temperatures are shown in blue. The forecasts are ten years long and launched every five years, and so the fan charts would overlap; to avoid this they are presented on two panels. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965).

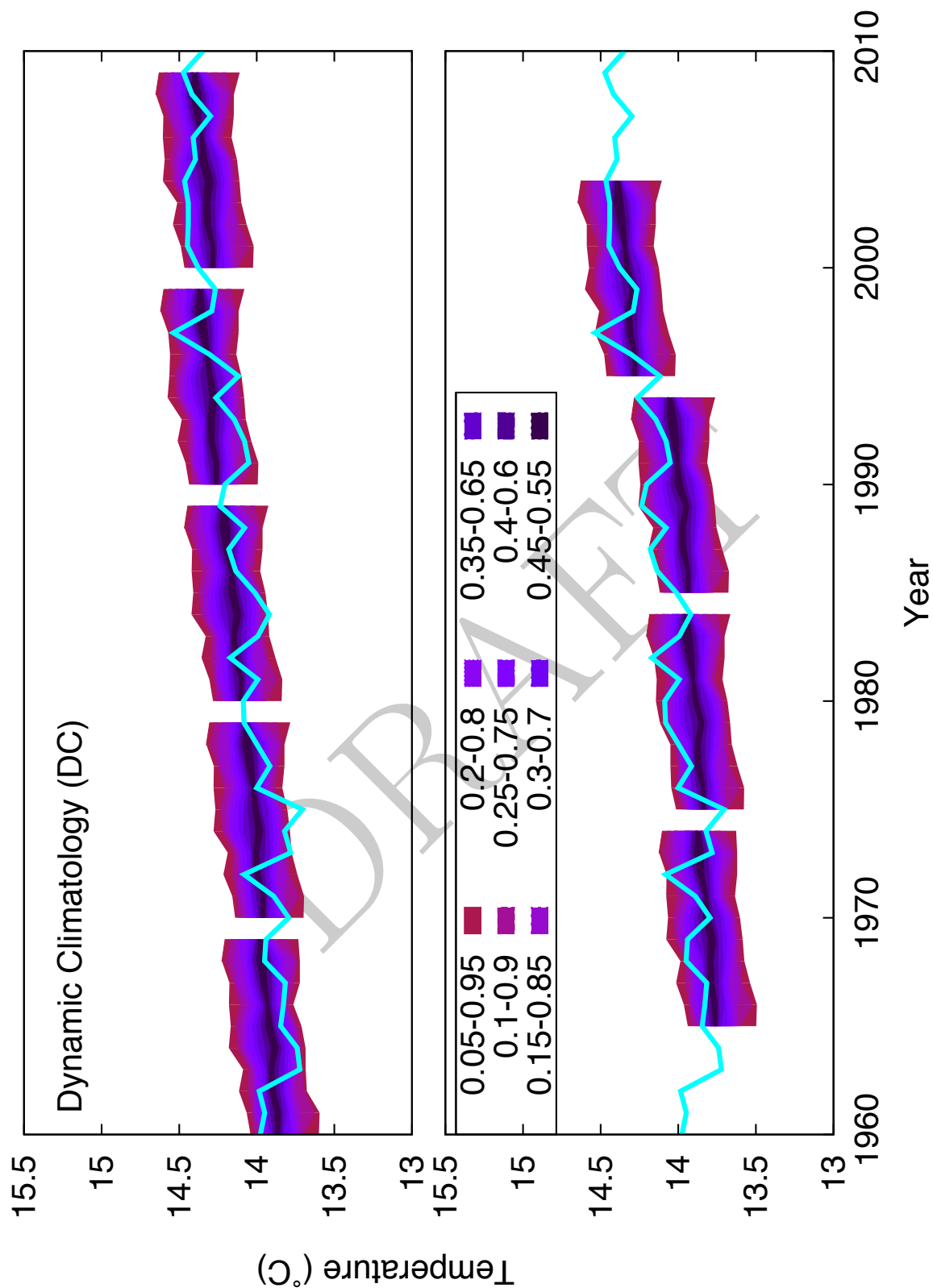


FIG. 5. Forecast distribution for every fifth launch from the Dynamic Climatology (DC) model for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The forecasts are ten years long and launched every five years, and so the fan charts would overlap; to avoid this they are presented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

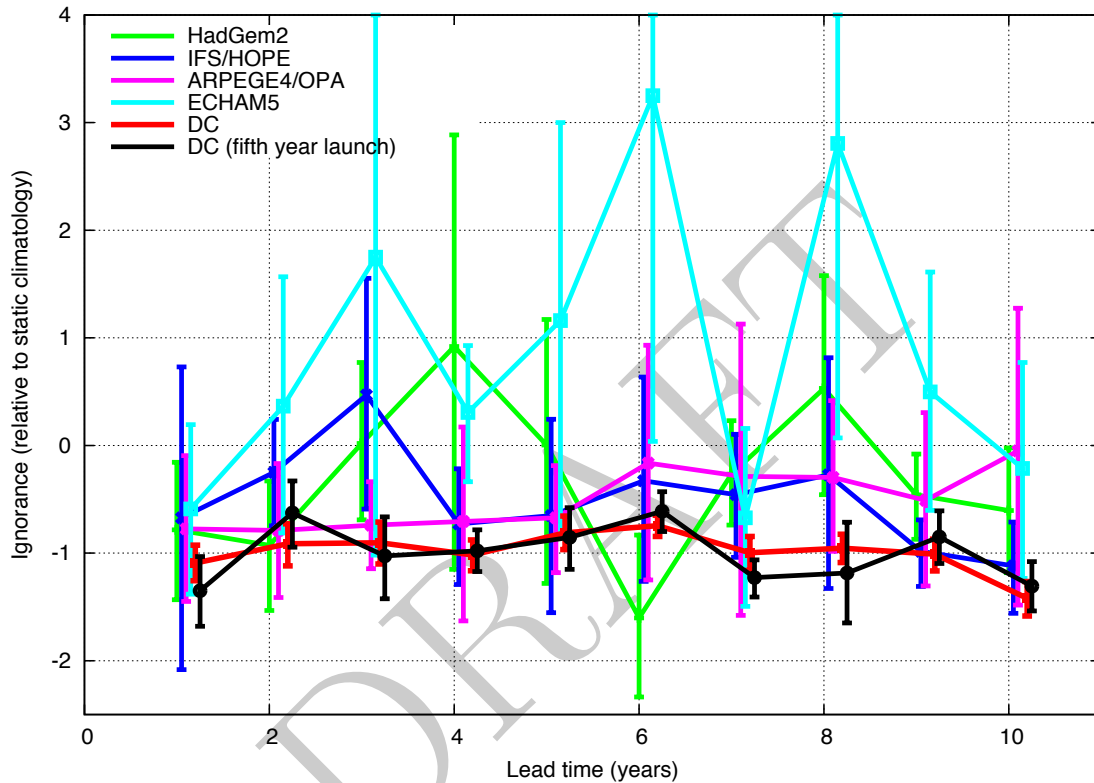


FIG. 6. Ignorance as a function of lead time for each of the four ENSEMBLES hindcast simulation models and the DC model relative to the static climatology. The bootstrap resampling intervals are illustrated at the 10-90th percent level. The DC model is shown to be significantly more skillful than static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

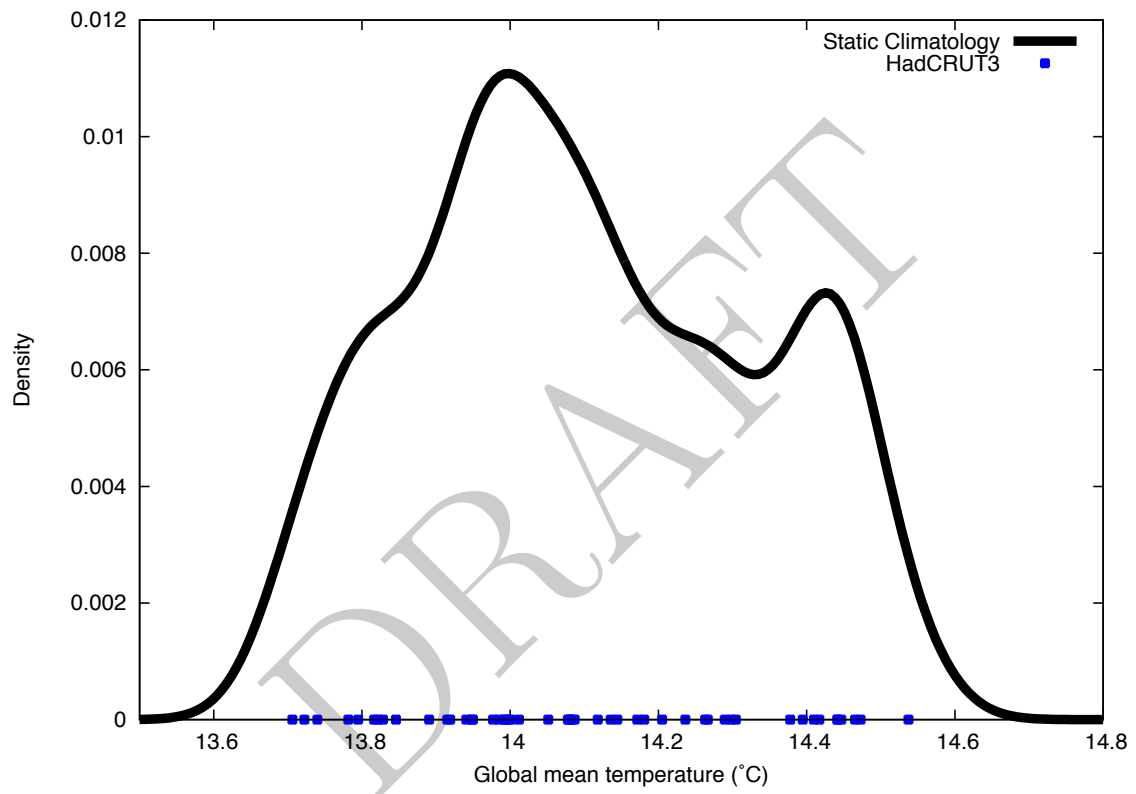


FIG. 7. Probability density for the static climatology used in the paper with observations over the period 1960-2010 (from HadCRUT3) illustrated as points on the x-axis for reference.

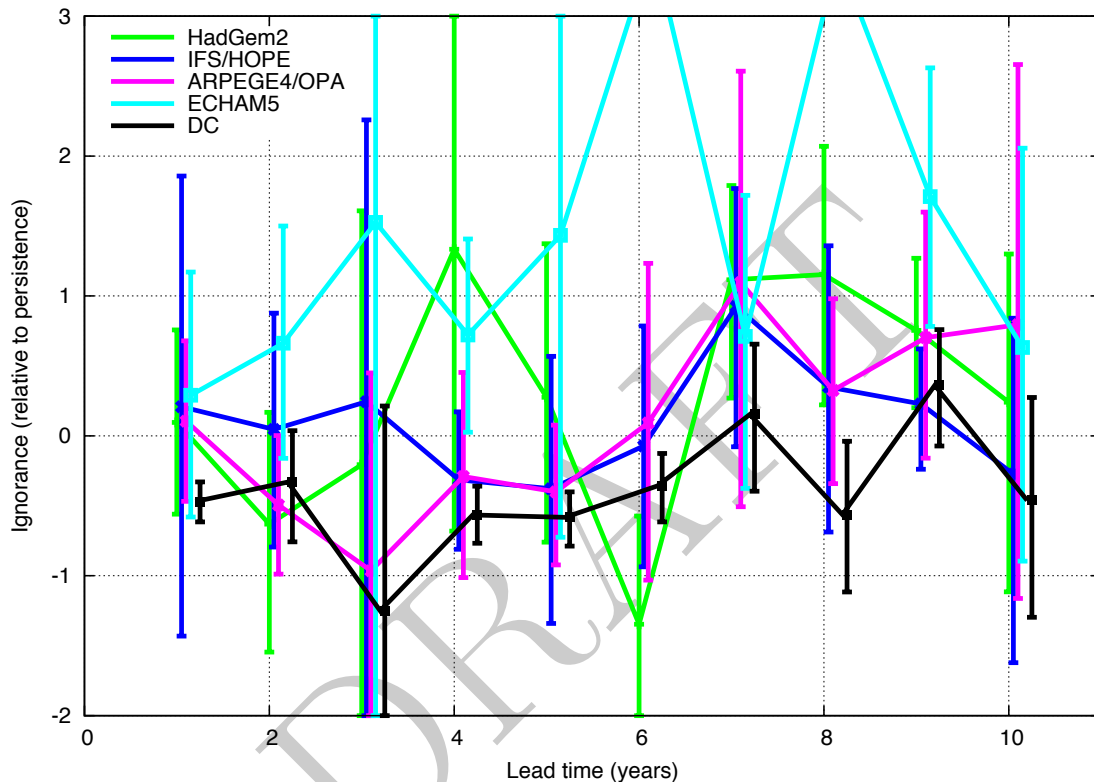


FIG. 8. Ignorance of the ENSEMBLES models and DC relative to persistence forecasts as a function of lead time. The DC model has negative relative Ignorance scores up to 6 years ahead, indicating it is significantly more skillful than persistence forecasts at early lead times. The ENSEMBLES models tend 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 10-90th percent level.

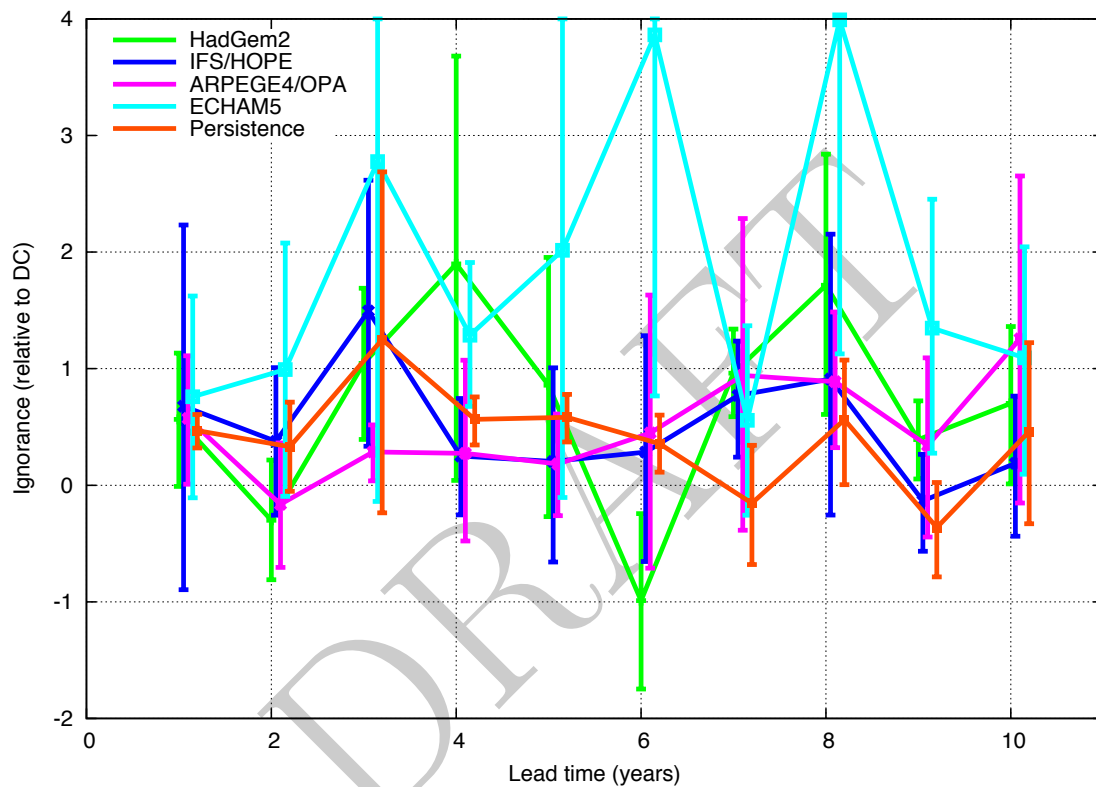


FIG. 9. Ignorance of the ENSEMBLES models relative to DC as a function of lead time. The bootstrap resampling intervals are illustrated at the 10-90th percent level. Note that the simulation models tend to have positive scores (less skill) than the DC model at every lead time.

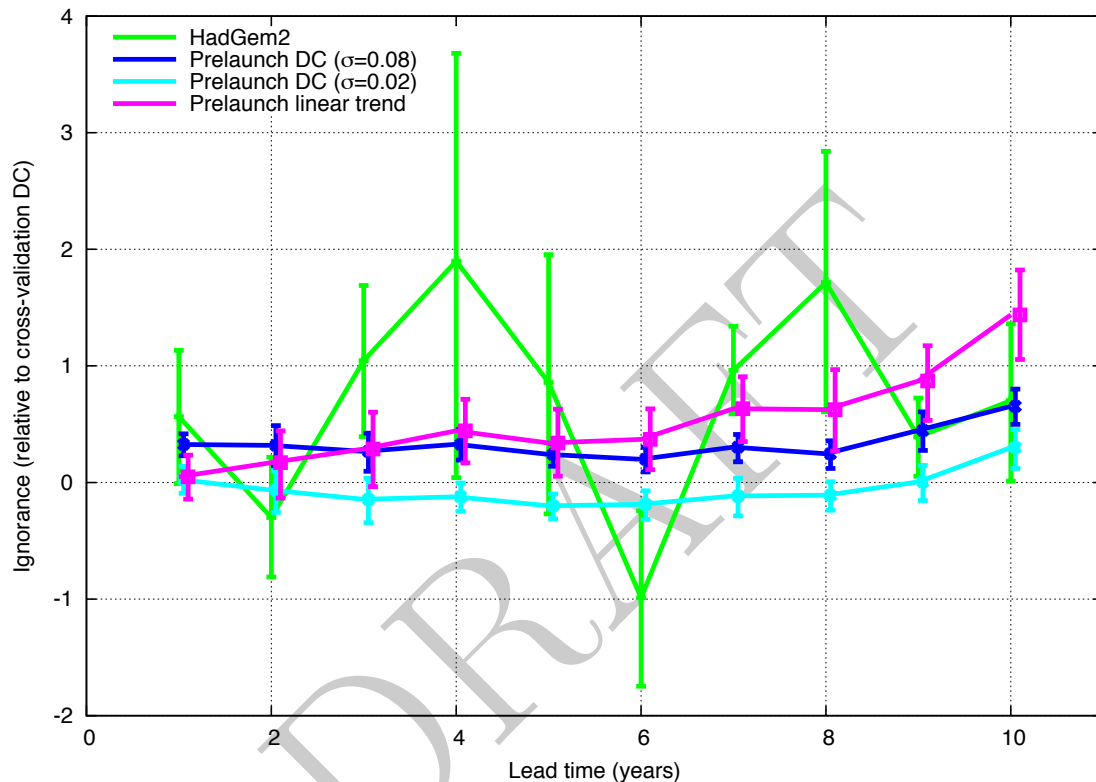


FIG. 10. Ignorance of the Prelaunch DC and Prelaunch Trend models relative to the standard DC model as a function of lead time. The HadGem2 model from ENSEMBLES is also shown. It is shown that the Prelaunch DC model is not significantly less skillful than the standard DC model and is robust to variations 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 10-90th percent level.

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

3 Material

4 Emma B. Suckling and Leonard A. Smith

5 October 21, 2013

6 **1. Introduction**

7 The following material is a supplement to ‘An evaluation of decadal probability forecasts
8 from state-of-the-art climate models’, in which the performance of simulation models from
9 Stream 2 of the ENSEMBLES decadal hindcasts (Doblas-Reyes et al. 2010) are contrasted
10 with the empirical dynamic climatology (DC) model over global and Giorgi region scales.
11 Further details about transforming ensemble simulations into probabilistic distributions are
12 presented below in Section 2. In Section 3 it is shown that the DC empirical model outper-
13 forms the ENSEMBLES simulation models by several bits at most lead times and for every
14 region studied. In Section 4 the robustness of the results in the main manuscript are evalu-
15 ated by using alternative proper scoring rules, namely the proper linear (PL) and continuous

16 ranked probability scores (CRPS). It is shown that the results are robust to the scoring rule
17 chosen. Finally, in Section 5 the performance of alternative empirical models are considered,
18 namely a ‘Prelaunch linear trend’ approach and ‘Prelaunch DC model’. It is shown that the
19 Prelaunch DC model performs to a similar quality as the standard DC approach employed
20 in the main manuscript, and is robust to the kernel parameters and anchor year chosen
21 to fit the model. Further details about generating the probabilistic DC forecasts and the
22 robustness of the results to the model parameter choices are also provided in Section 5.

23 **2. Probabilistic forecast distributions for the ENSEM-** 24 **BLES simulation models**

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

29 Information contamination is a significant concern in the evaluation of decadal forecasts.
30 Given that the total duration of hindcast experiments is typically fifty years, there are
31 very few independent decadal periods in the forecast-outcome archive. Cross-validation ap-
32 proaches attempt to maximise the size of the forecast-outcome archive (to increase statistical
33 significance) while avoiding the use of information from a given forecast target period being
34 used in the evaluation of that forecast. It is crucial to also avoid information contamination
35 by inadvertently using information from the target decade when interpreting the ensemble

36 into a forecast distribution (Bröcker and Smith 2008). This cannot be done rigorously in
37 the case of simulation models, as the structure and parameters of the models themselves
38 have evolved in light of the observations of the last fifty years. The true-leave-one-out cross-
39 validation procedure described in the main manuscript avoids any explicit use of data from
40 within the target forecast period, even as its implicit use cannot be avoided. In practice this
41 is achieved by leaving out the target decade, then using a standard leave-one-out procedure
42 to fit the kernel parameters for each forecast in turn.

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

54 The bottom panels of figure 4 show the mean Ignorance as a function of kernel offset
55 over the full set of hindcast simulations. Once again the vertical bars indicate the values of
56 offset that were used for the individual forecasts, based on minimising Ignorance through the
57 true-leave-one out protocol. Once again, at lead time six the fitting protocol favours a kernel
58 offset under true-leave-one-out cross-validation that falls outside the minimum Ignorance

59 value without cross-validation. The result for lead time one is typical of the kernel offset
60 values attained for the other lead times.

61 **3. Regional analysis**

62 Figures 5 to 25 show Ignorance as a function of lead time for each of the ENSEMBLES
63 models relative to the DC empirical model for surface air temperature over each of the
64 land-based Giorgi regions (Giorgi 2002). At Giorgi region scales the decadal probability
65 forecasts from the ENSEMBLES models perform to a similar quality as for the global mean
66 temperature in some cases, or significantly worse in others. In some regions and at some
67 lead times DC outperforms the ENSEMBLES models by more than 4 bits; DC placing over
68 $16 (2^4)$ times more probability mass on the verification than the simulation model. In these
69 figures no simulation model demonstrates skill significantly above the DC model for any
70 lead time or any region; positive values of the relative Ignorance performance measure are
71 reported in all of the cases below.

72 **4. Robustness to the performance measure**

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

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

82 **5. Alternative empirical models**

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

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

99 available before the forecast launch time are used.

100 Examining the of the score to variations in the parameters can reveal overfitting. Figure
101 28 shows the skill of the Prelaunch DC for values of the kernel width ranging from 0.02 to
102 0.16 for forecast lead times of one to ten years. Ignorance relative to the standard DC model
103 is shown. The sensitivity of the Prelaunch DC model to variation in the starting date for the
104 forecast-outcome archive (not shown) is less than the sensitivity to the kernel width. Start
105 dates from 1900 to 1950 were considered; the later start dates tend to yield more skilful
106 models. The Prelaunch DC discussed in the main text uses a start date of 1950 and a width
107 of 0.08, although this value does not correspond to the lowest in-sample skill - as shown in
108 figure 29. Furthermore the ensemble interpretation of the simulations models reported in
109 this paper use data both before and after the target window, giving those simulation models
110 an unquantified advantage over the empirical models defined here.

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

122 Prelaunch DC model parameters.

123 A Prelaunch trend model is also discussed in the main text. This model is fully defined
124 by the initial time anchor from which the trend is estimated. Figure 30 shows the skill of
125 this model relative to the standard DC model for several anchor times between 1900 and
126 1950. The results in the main text use the 1950 anchor time. It is shown that although
127 there is some sensitivity to the anchor time, all the Prelaunch trend models are generally
128 less skillful than the standard DC model.

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

133

134

REFERENCES

135 Bröcker, J. and L. A. Smith, 2008: From ensemble forecasts to predictive distributions.

136 *Tellus A*, **60** (4), 663-678.

137 Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Fore-

138 cast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. *Technical*

139 *Memorandum ECMWF*, **621**.

- 140 Giorgi, F., 2002: Variability and trends of sub-continental scale surface climate in the twen-
141 tieth century. part i: observations. *Climate Dynamics*, **18**, 675–691.
- 142 Good, I. J., 1952: Rational decisions. *Journal of the Royal Statistical Society*, **XIV (1)**,
143 107–114.
- 144 Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast verification: A practitioner's guide in*
145 *atmospheric science*. John Wiley and Sons Ltd.

DRAFT

List of Figures

- 146 **List of Figures**
- 147 1 Forecast distributions for IFS/HOPE (ECMWF) for the 5-95th percentile.
148 The HadCRUT3 observed temperatures are shown in blue. Each forecast is
149 ten years long and they are launched every five years. To avoid overlap of
150 the fan charts they are presented on two panels. The top (bottom) panel
151 illustrates forecasts launched in ten year intervals from 1960 (1965). It is
152 shown that the observed global mean temperature often falls outside the 5-
153 95th percentile of the predicted distributions. 16
- 154 2 Forecast distributions for ARPEGE/OPA (CERFACS) for the 5-95th per-
155 centile. The HadCRUT3 observed temperatures are shown in blue. The top
156 (bottom) panel illustrates forecasts launched in ten year intervals from 1960
157 (1965). It is shown that the observed global mean temperature often falls
158 outside the 5-95th percentile of the predicted distributions. 17
- 159 3 Forecast distributions for ECHAM5 (IFM-GEOMAR) for the 5-95th per-
160 centile. The HadCRUT3 observed temperatures are shown in blue. The top
161 (bottom) panel illustrates forecasts launched in ten year intervals from 1960
162 (1965). It is shown that the observed global mean temperature falls outside
163 the 5-95th percentile of the predicted distributions on several occasions. 18

- 164 4 Ignorance as a function of kernel dressing parameters over the full set of
165 hindcast simulations (*i.e.* with no cross-validation) for the HadGem2 model
166 at lead time one (a and c) and lead time six (b and d). The top panels (a and
167 b) show the score as a function of the kernel width parameter and the bottom
168 panels (c and d) show the score as a function of the kernel offset parameter.
169 The vertical bars in each case illustrate the kernel parameters obtained for
170 each individual forecast under true-leave-one-out cross-validation. That there
171 are fewer than nine vertical bars indicates that the kernel parameter values
172 shown were obtained for several forecasts in the set. Results for lead times
173 two to five and seven to ten (not shown) are similar to those shown for lead
174 time one. 19
- 175 5 Ignorance of the ENSEMBLES simulation models relative to the DC model for
176 Alaska. Scores above zero indicate that the DC model outperforms the simu-
177 lation models, placing significantly more probability on the observed outcome
178 than the ENSEMBLES models. 20
- 179 6 Ignorance of the ENSEMBLES simulation models relative to the DC model
180 for Amazon Basin. Scores above zero indicate that the DC model outperforms
181 the simulation models, placing significantly more probability on the observed
182 outcome than the ENSEMBLES models. 21
- 183 7 Ignorance of the ENSEMBLES simulation models relative to the DC model
184 for Australia. Scores above zero indicate that the DC model outperforms
185 the simulation models, placing significantly more probability on the observed
186 outcome than the ENSEMBLES models. 22

- 187 8 Ignorance of the ENSEMBLES simulation models relative to the DC model for
188 Central America. Scores above zero indicate that the DC model outperforms
189 the simulation models, placing significantly more probability on the observed
190 outcome than the ENSEMBLES models. 23
- 191 9 Ignorance of the ENSEMBLES simulation models relative to the DC model
192 for Central Asia. Scores above zero indicate that the DC model outperforms
193 the simulation models, placing significantly more probability on the observed
194 outcome than the ENSEMBLES models. 24
- 195 10 Ignorance of the ENSEMBLES simulation models relative to the DC model
196 for Central North America. Scores above zero indicate that the DC model
197 outperforms the simulation models, placing significantly more probability on
198 the observed outcome than the ENSEMBLES models. 25
- 199 11 Ignorance of the ENSEMBLES simulation models relative to the DC model
200 for Eastern Africa. Scores above zero indicate that the DC model outperforms
201 the simulation models, placing significantly more probability on the observed
202 outcome than the ENSEMBLES models. 26
- 203 12 Ignorance of the ENSEMBLES simulation models relative to the DC model
204 for Eastern North America. Scores above zero indicate that the DC model
205 outperforms the simulation models, placing significantly more probability on
206 the observed outcome than the ENSEMBLES models. 27

- 207 13 Ignorance of the ENSEMBLES simulation models relative to the DC model
208 for East Asia. Scores above zero indicate that the DC model outperforms
209 the simulation models, placing significantly more probability on the observed
210 outcome than the ENSEMBLES models. 28
- 211 14 Ignorance of the ENSEMBLES simulation models relative to the DC model
212 for Greenland. Scores above zero indicate that the DC model outperforms
213 the simulation models, placing significantly more probability on the observed
214 outcome than the ENSEMBLES models. 29
- 215 15 Ignorance of the ENSEMBLES simulation models relative to the DC model
216 for Mediterranean Basin. Scores above zero indicate that the DC model out-
217 performs the simulation models, placing significantly more probability on the
218 observed outcome than the ENSEMBLES models. 30
- 219 16 Ignorance of the ENSEMBLES simulation models relative to the DC model
220 for North Asia. Scores above zero indicate that the DC model outperforms
221 the simulation models, placing significantly more probability on the observed
222 outcome than the ENSEMBLES models. 31
- 223 17 Ignorance of the ENSEMBLES simulation models relative to the DC model for
224 Northern Europe. Scores above zero indicate that the DC model outperforms
225 the simulation models, placing significantly more probability on the observed
226 outcome than the ENSEMBLES models. 32

- 227 18 Ignorance of the ENSEMBLES simulation models relative to the DC model for
228 Southern Africa. Scores above zero indicate that the DC model outperforms
229 the simulation models, placing significantly more probability on the observed
230 outcome than the ENSEMBLES models. 33
- 231 19 Ignorance of the ENSEMBLES simulation models relative to the DC model for
232 Sahara. Scores above zero indicate that the DC model outperforms the simu-
233 lation models, placing significantly more probability on the observed outcome
234 than the ENSEMBLES models. 34
- 235 20 Ignorance of the ENSEMBLES simulation models relative to the DC model
236 for South Asia. Scores above zero indicate that the DC model outperforms
237 the simulation models, placing significantly more probability on the observed
238 outcome than the ENSEMBLES models. 35
- 239 21 Ignorance of the ENSEMBLES simulation models relative to the DC model
240 for Southeast Asia. Scores above zero indicate that the DC model outperforms
241 the simulation models, placing significantly more probability on the observed
242 outcome than the ENSEMBLES models. 36
- 243 22 Ignorance of the ENSEMBLES simulation models relative to the DC model
244 for Southern South America. Scores above zero indicate that the DC model
245 outperforms the simulation models, placing significantly more probability on
246 the observed outcome than the ENSEMBLES models. 37

247	23	Ignorance of the ENSEMBLES simulation models relative to the DC model for	
248		Tibet. Scores above zero indicate that the DC model outperforms the simu-	
249		lation models, placing significantly more probability on the observed outcome	
250		than the ENSEMBLES models.	38
251	24	Ignorance of the ENSEMBLES simulation models relative to the DC model	
252		for Western Africa. Scores above zero indicate that the DC model outperforms	
253		the simulation models, placing significantly more probability on the observed	
254		outcome than the ENSEMBLES models.	39
255	25	Ignorance of the ENSEMBLES simulation models relative to the DC model	
256		for Western North America. Scores above zero indicate that the DC model	
257		outperforms the simulation models, placing significantly more probability on	
258		the observed outcome than the ENSEMBLES models.	40
259	26	Proper linear score for each of the ENSEMBLES simulation models and the	
260		DC empirical model. Lower scores indicate better foecasts. The DC model is	
261		shown to outperform the simulations models at most lead times.	41
262	27	CRPS score for each of the ENSEMBLES simulation models and the DC	
263		empirical model. Lower scores indicate better forecasts. The DC model is	
264		shown to outperform the simulations models at most lead times.	42
265	28	Ignorance of the Prelaunch DC empirical model with kernel widths as la-	
266		belled relative to the cross-validation DC model. Increasing the kernel width	
267		parameter from 0.02 to 0.16 results in a loss of skill of approximately half	
268		a bit, although for the kernel width value used in this paper (0.08) there is	
269		shown to be no significant loss of skill relative to the standard DC model.	43

270 29 Ignorance as a function of the kernel width parameter over the full set of
271 hindcast simulations (*i.e.* with no cross-validation) for the DC (left panels)
272 and Prelaunch DC (right panels) models at lead time one (a and b), six
273 (c and d) and ten (e and f). The vertical bars in each case illustrate the
274 kernel width parameters employed in the main manuscript. In the DC model
275 parameters were attained through true-leave-one-out cross-validation. In the
276 Prelaunch DC model a kernel spread value of 0.08 was chosen for comparison
277 with DC and to test the robustness of the results to choices in the parameters
278 for ensemble interpretation (although this value does not correspond to the
279 lowest value of in-sample skill). 44

280 30 Ignorance of the Prelaunch trend empirical model for different anchor times
281 relative to the cross-validation DC model. Scores above zero indicate that
282 DC outperforms the Prelaunch Trend model by up to half a bit at early lead
283 times, and up to two bits (DC placing up to 4 times more probability on the
284 observed outcome than the Prelaunch Trend model) up to ten years ahead,
285 depending on the anchor year for the trend model. 45

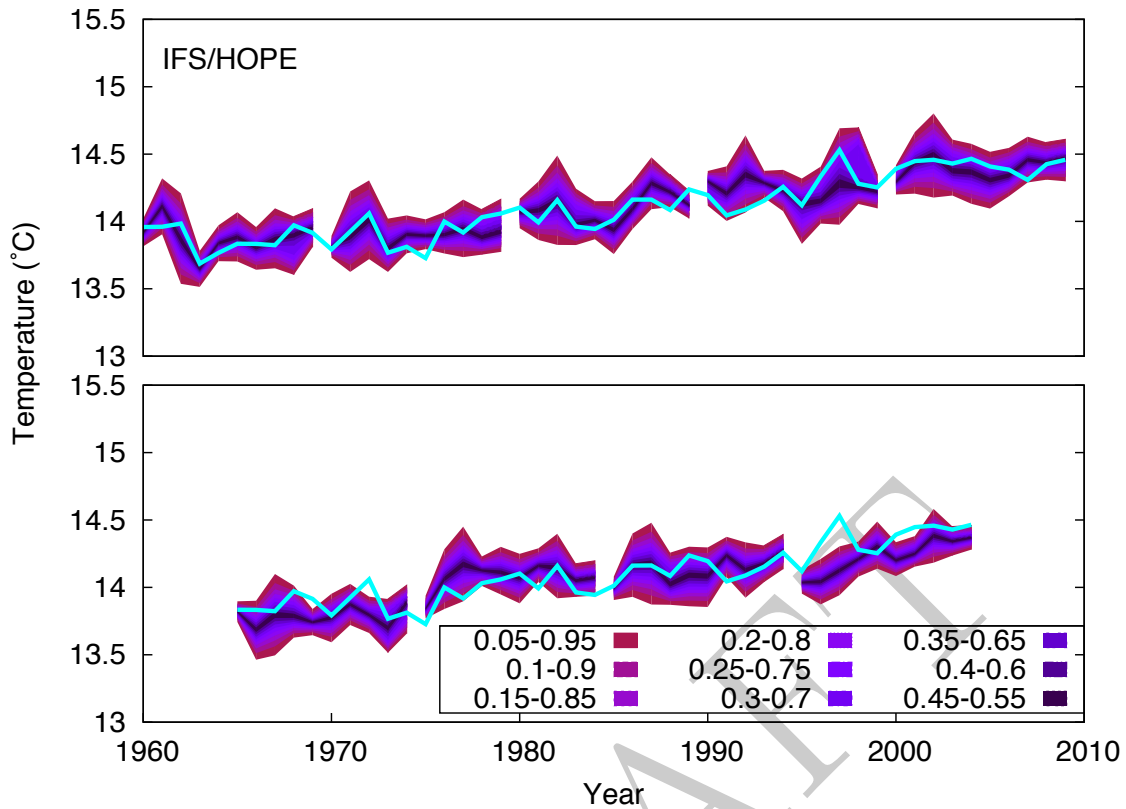


FIG. 1. Forecast distributions for IFS/HOPE (ECMWF) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. Each forecast is ten years long and they are launched every five years. To avoid overlap of the fan charts they are presented on two panels. The top (bottom) panel illustrates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

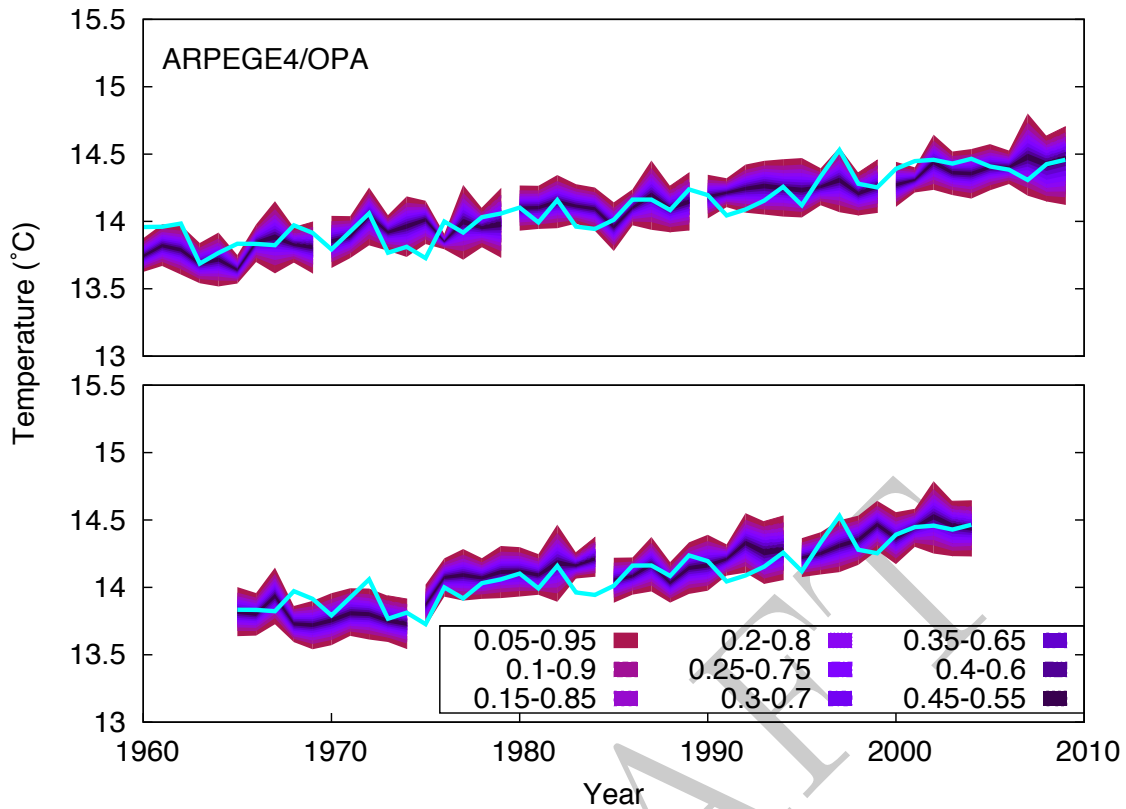


FIG. 2. Forecast distributions for ARPEGE/OPA (CERFACS) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature often falls outside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

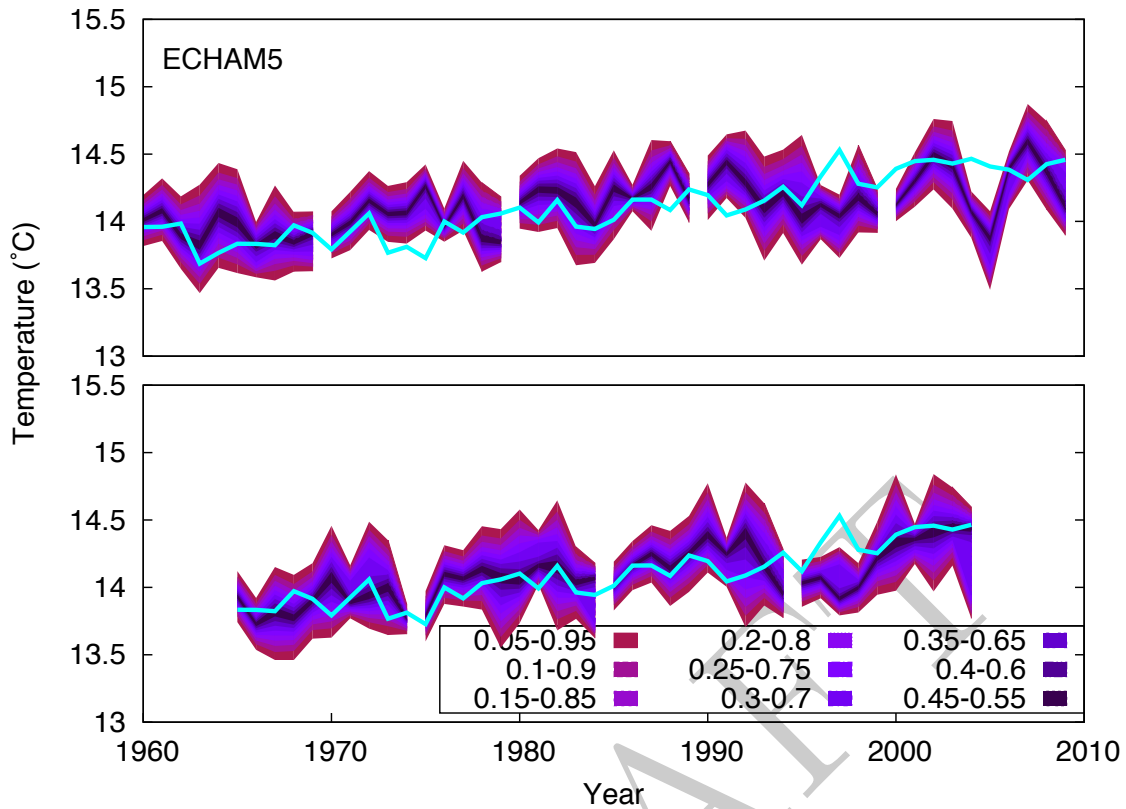


FIG. 3. Forecast distributions for ECHAM5 (IFM-GEOMAR) for the 5-95th percentile. The HadCRUT3 observed temperatures are shown in blue. The top (bottom) panel illustrates forecasts launched in ten year intervals from 1960 (1965). It is shown that the observed global mean temperature falls outside the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

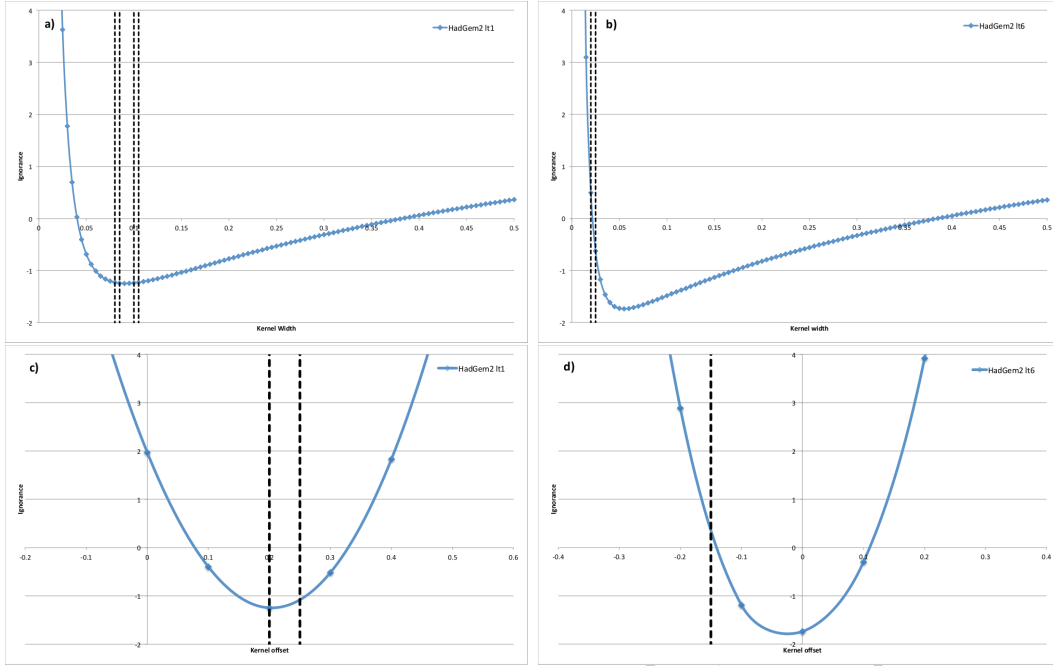


FIG. 4. Ignorance as a function of kernel dressing parameters over the full set of hindcast simulations (*i.e.* with no cross-validation) for the HadGem2 model at lead time one (a and c) and lead time six (b and d). The top panels (a and b) show the score as a function of the kernel width parameter and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

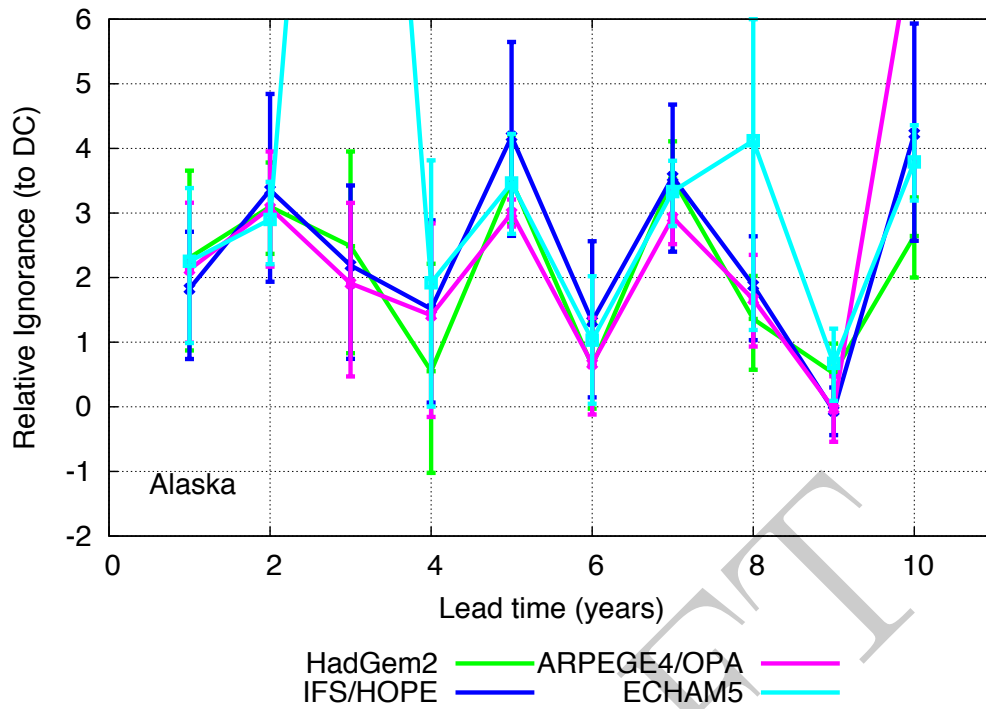


FIG. 5. Ignorance of the ENSEMBLES simulation models relative to the DC model for Alaska. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

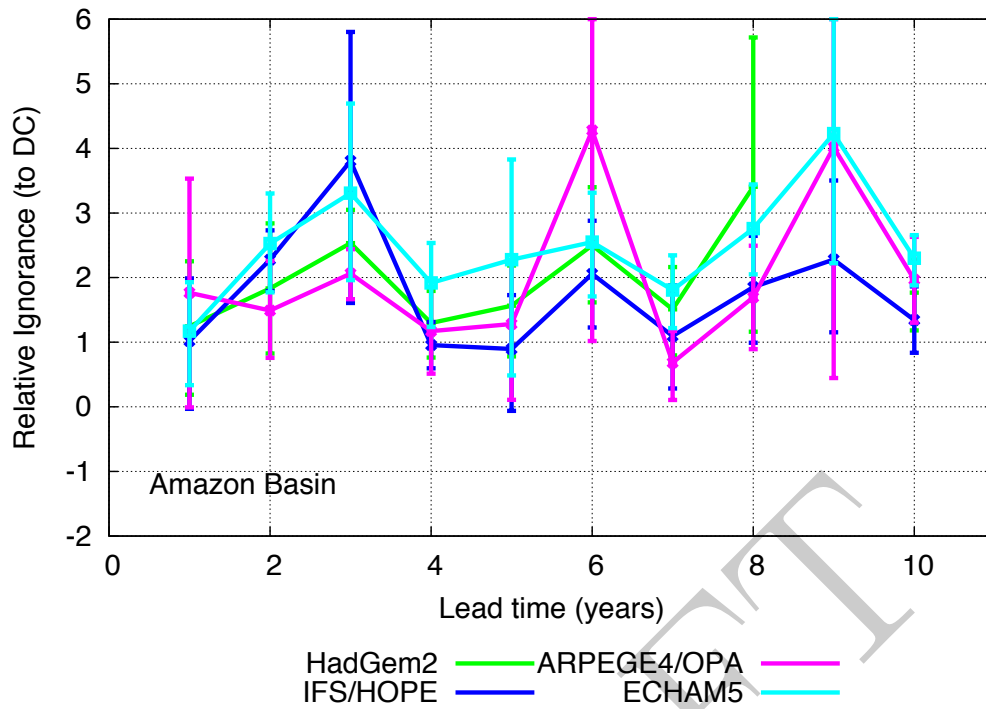


FIG. 6. Ignorance of the ENSEMBLES simulation models relative to the DC model for Amazon Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

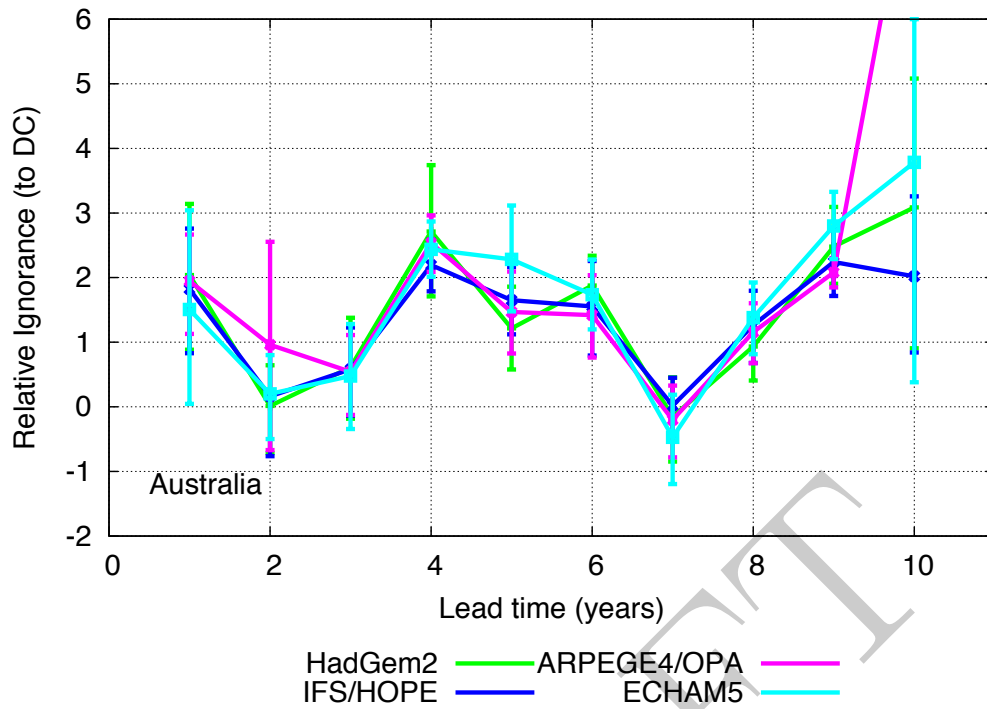


FIG. 7. Ignorance of the ENSEMBLES simulation models relative to the DC model for Australia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

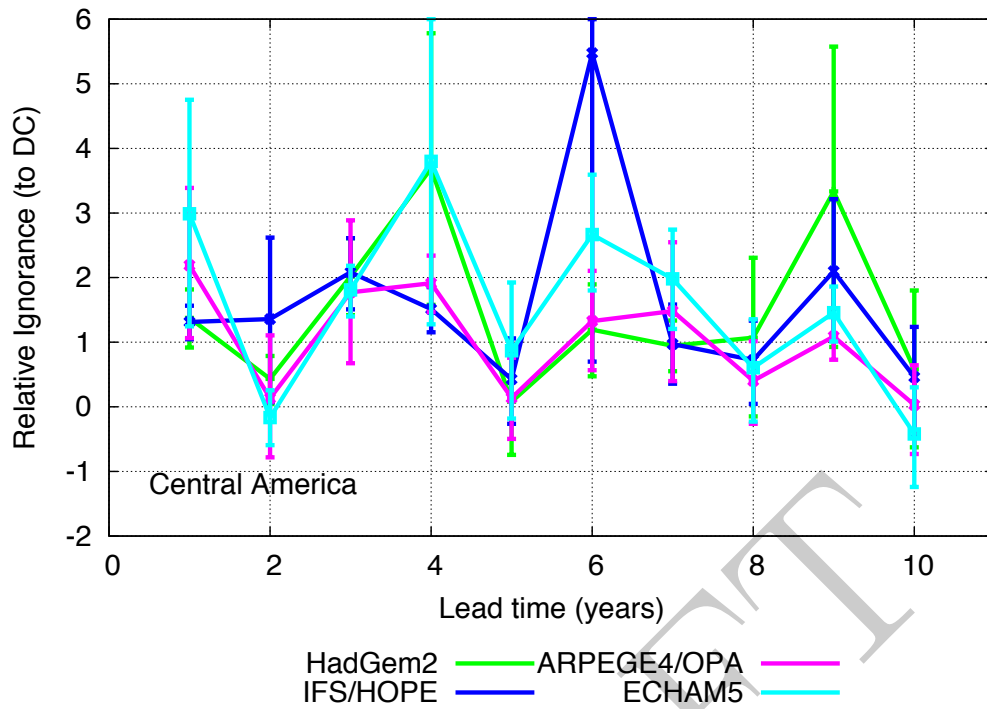


FIG. 8. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

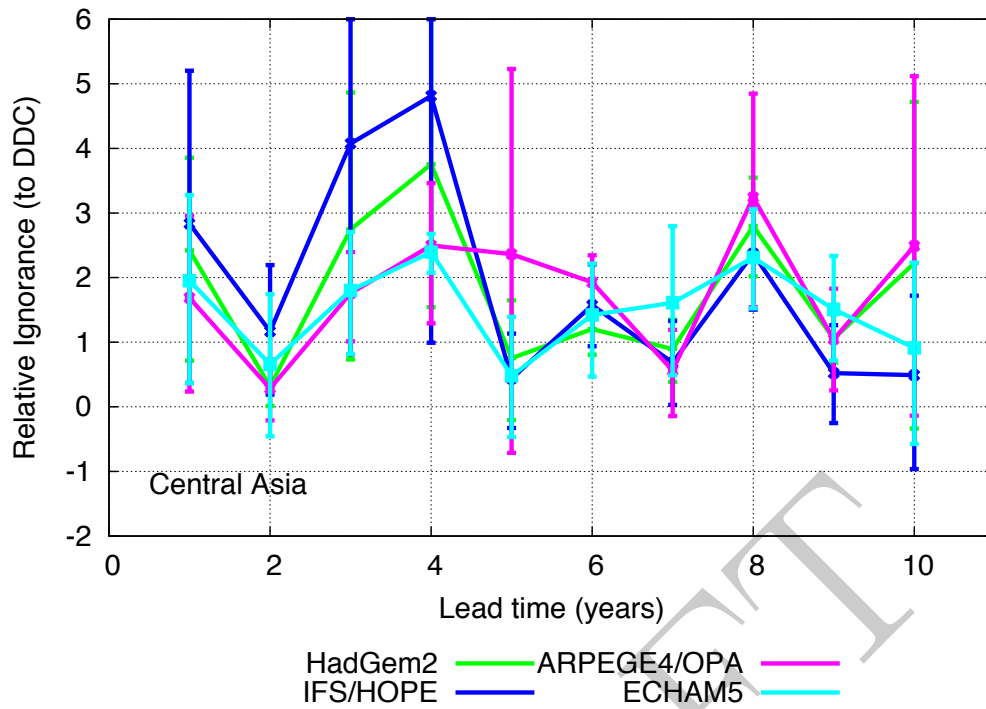


FIG. 9. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

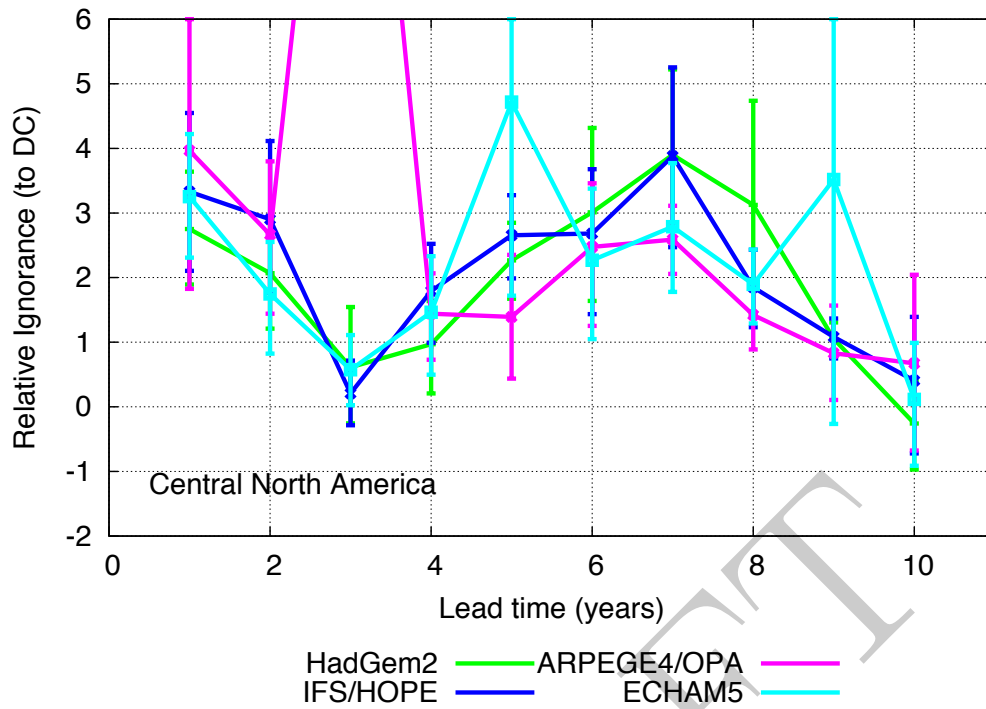


FIG. 10. Ignorance of the ENSEMBLES simulation models relative to the DC model for Central North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

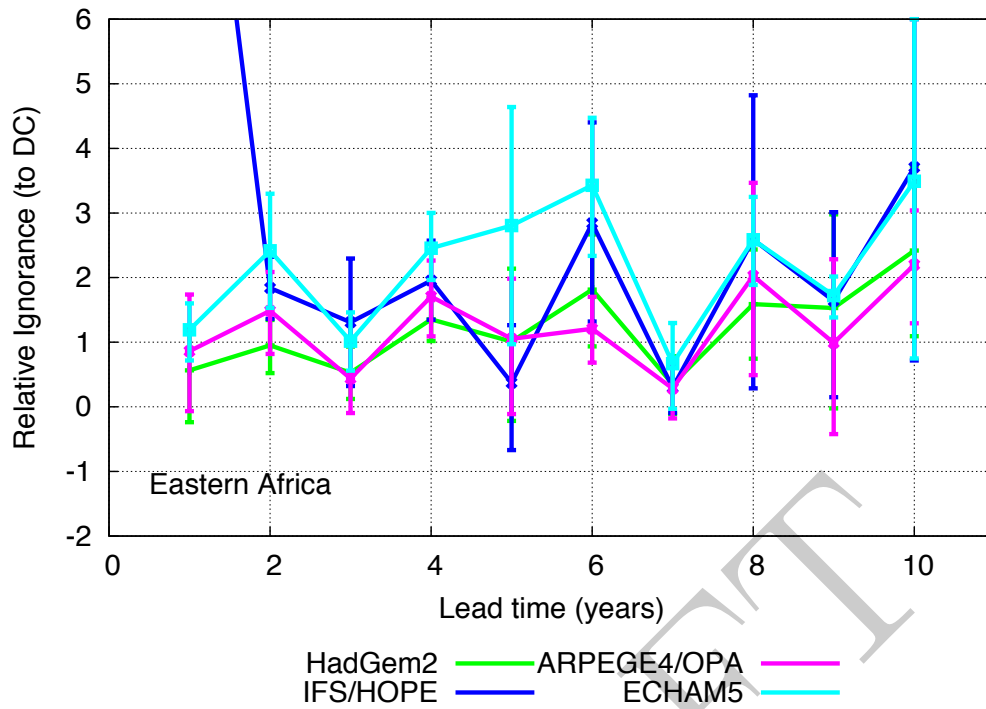


FIG. 11. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

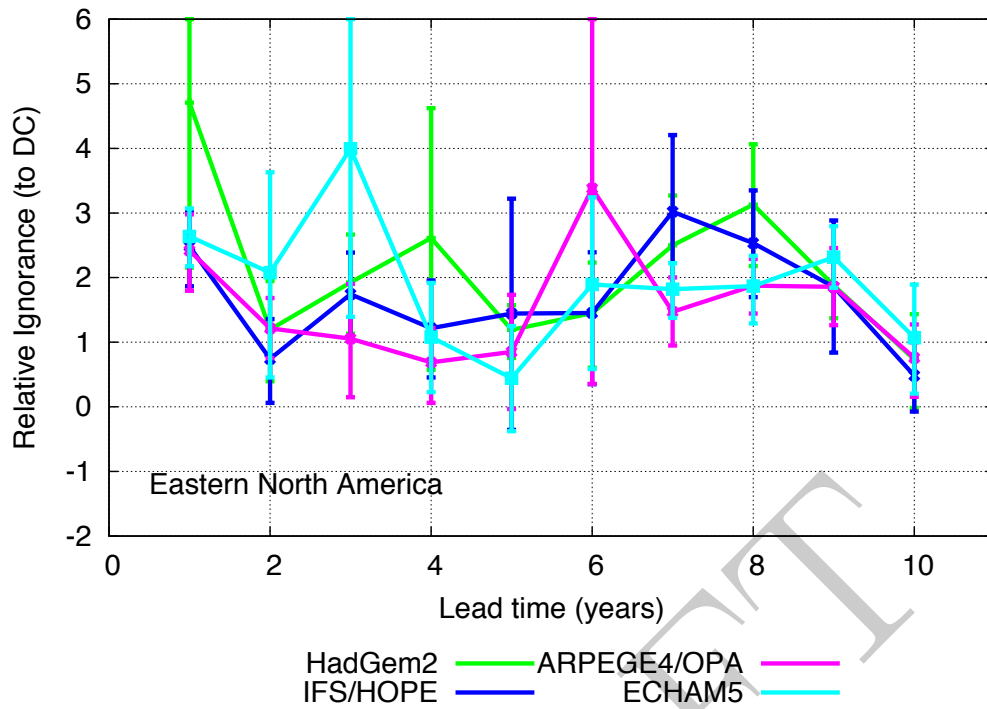


FIG. 12. Ignorance of the ENSEMBLES simulation models relative to the DC model for Eastern North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

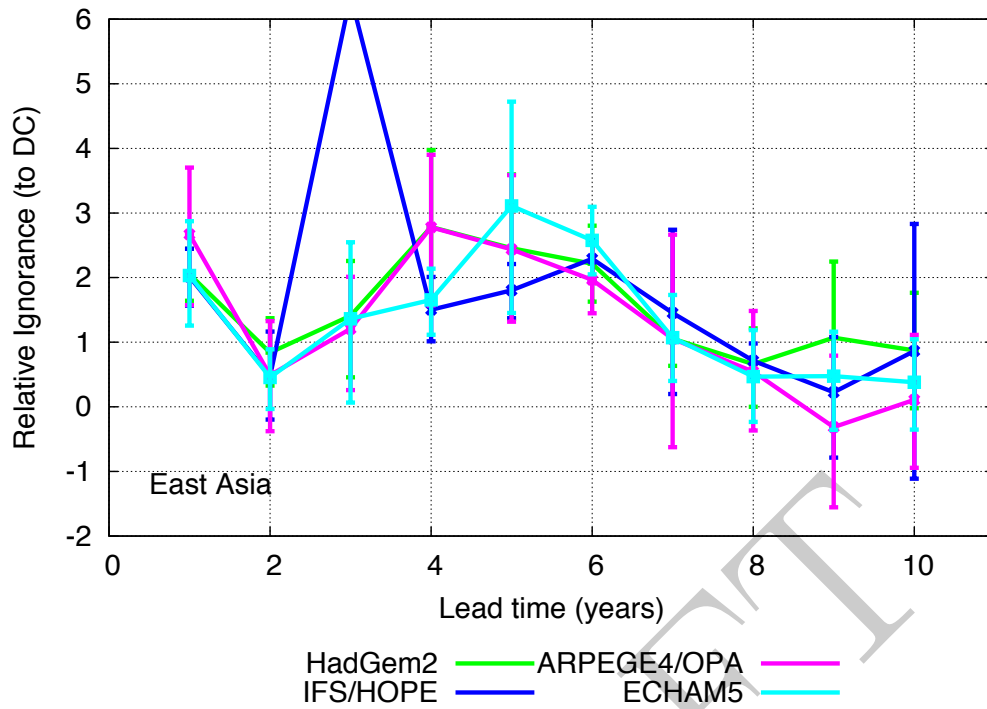


FIG. 13. Ignorance of the ENSEMBLES simulation models relative to the DC model for East Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

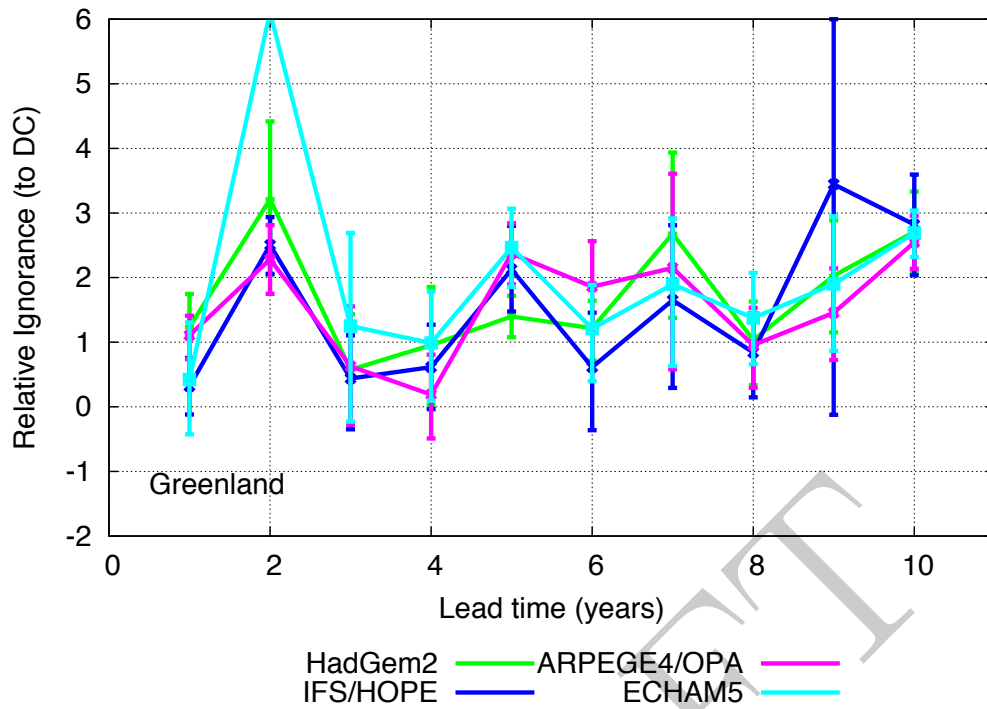


FIG. 14. Ignorance of the ENSEMBLES simulation models relative to the DC model for Greenland. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

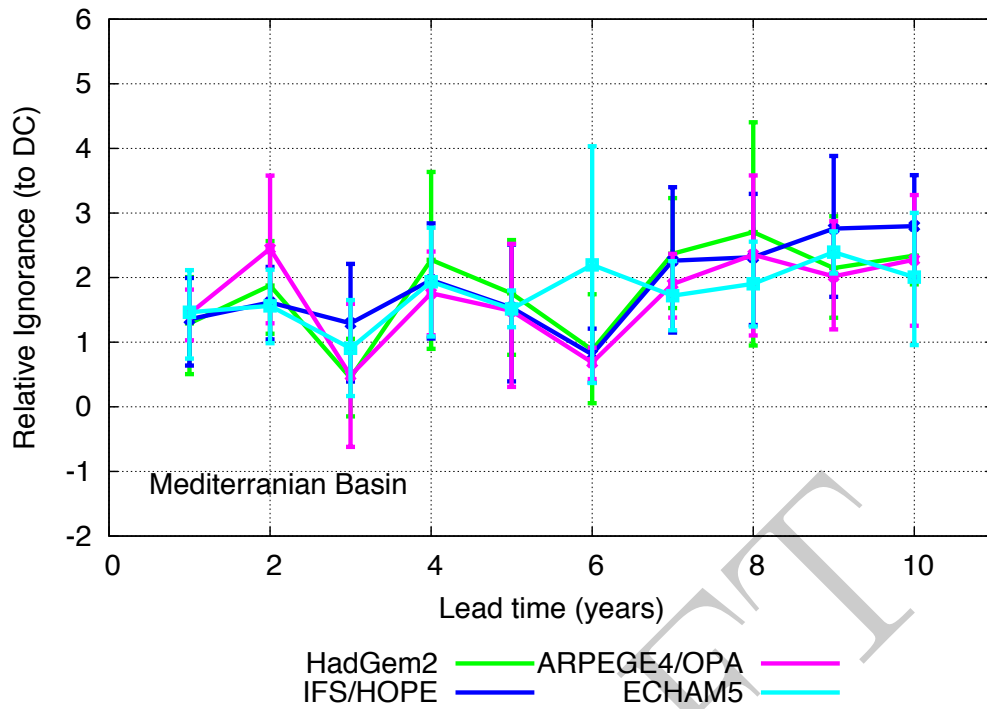


FIG. 15. Ignorance of the ENSEMBLES simulation models relative to the DC model for Mediterranean Basin. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

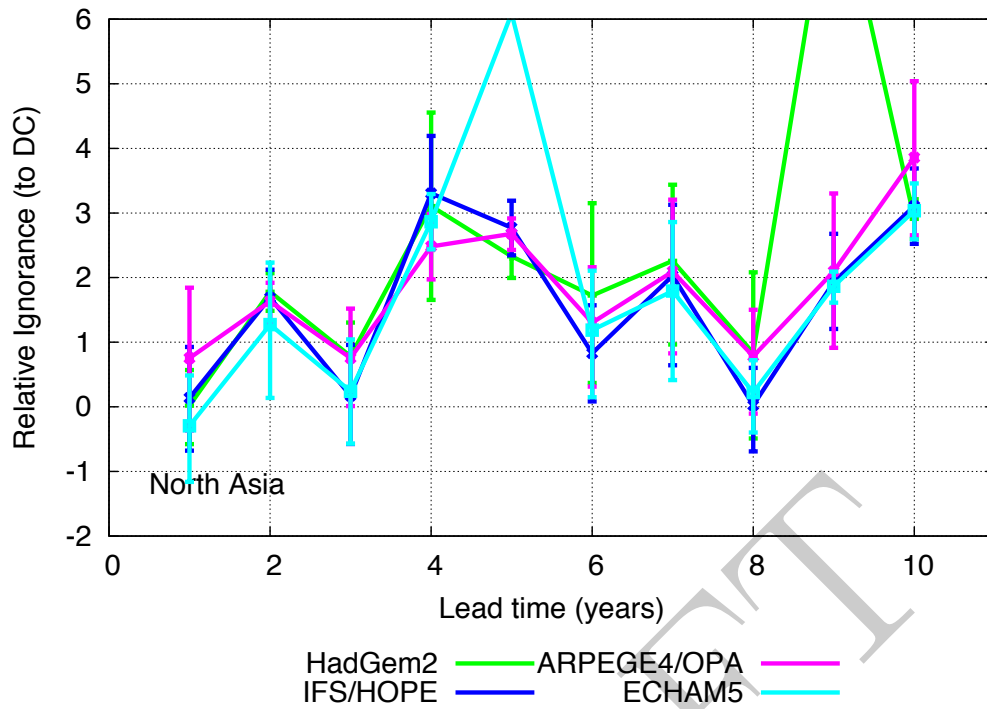


FIG. 16. Ignorance of the ENSEMBLES simulation models relative to the DC model for North Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

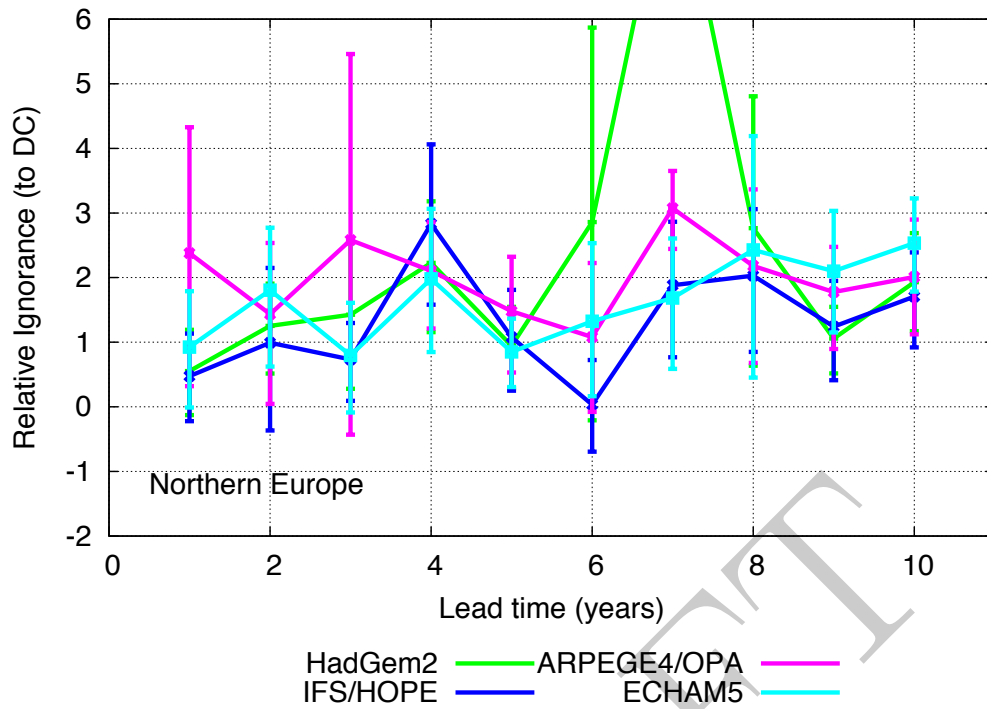


FIG. 17. Ignorance of the ENSEMBLES simulation models relative to the DC model for Northern Europe. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

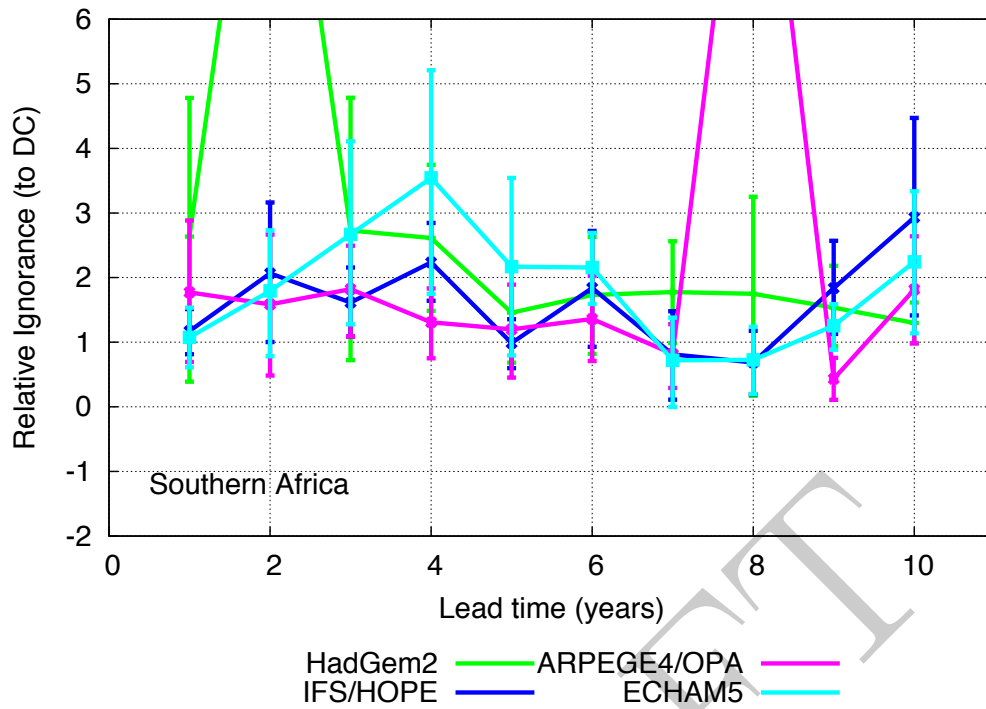


FIG. 18. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

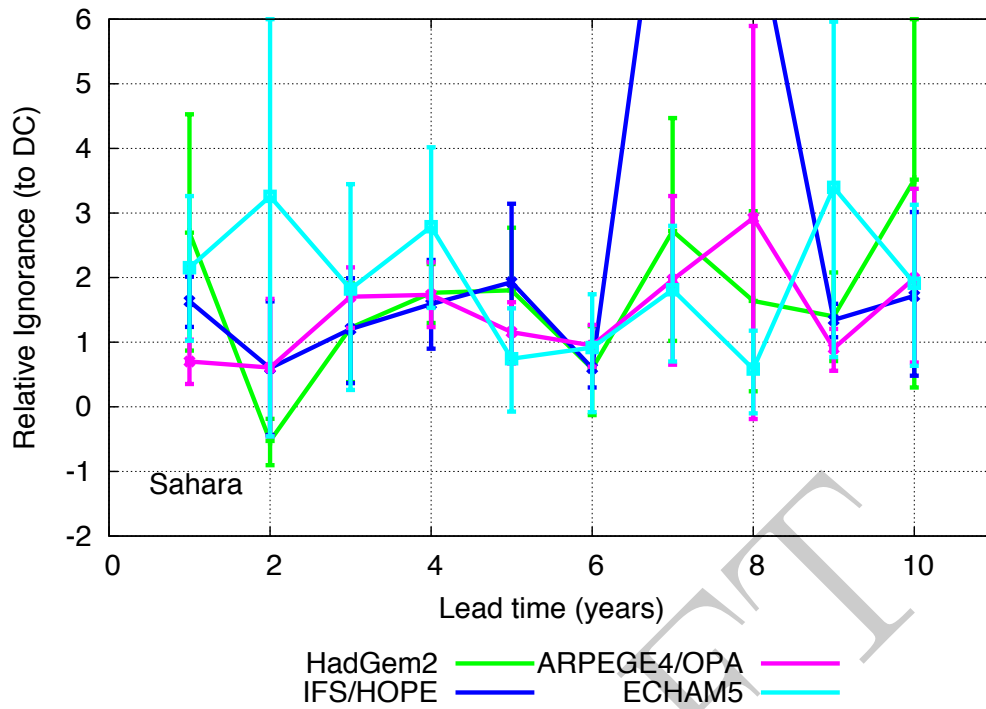


FIG. 19. Ignorance of the ENSEMBLES simulation models relative to the DC model for Sahara. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

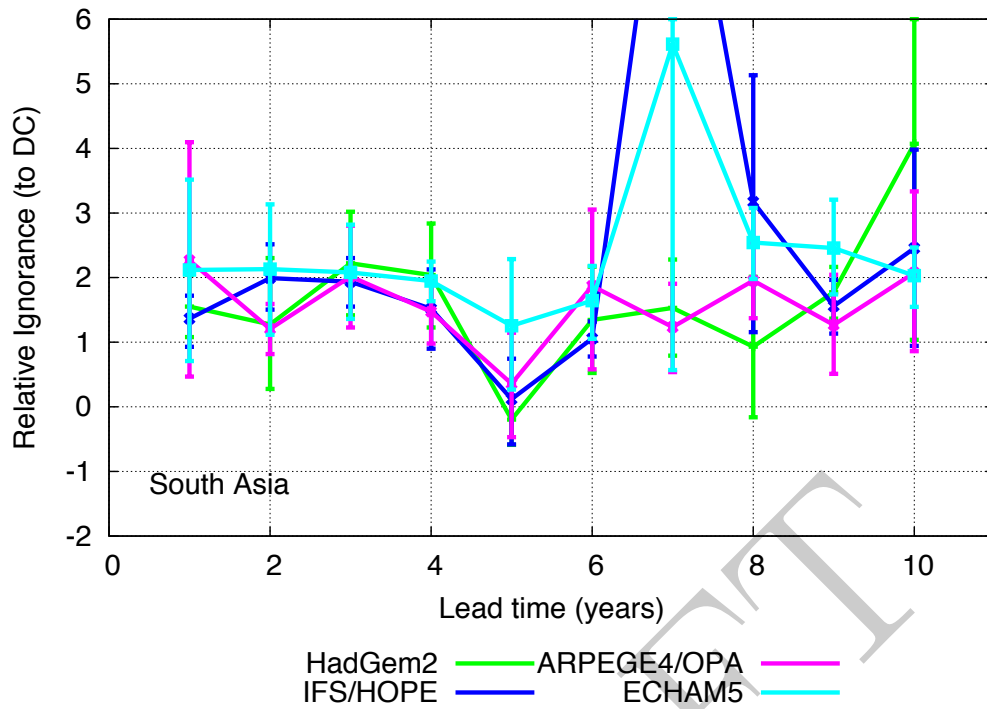


FIG. 20. Ignorance of the ENSEMBLES simulation models relative to the DC model for South Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

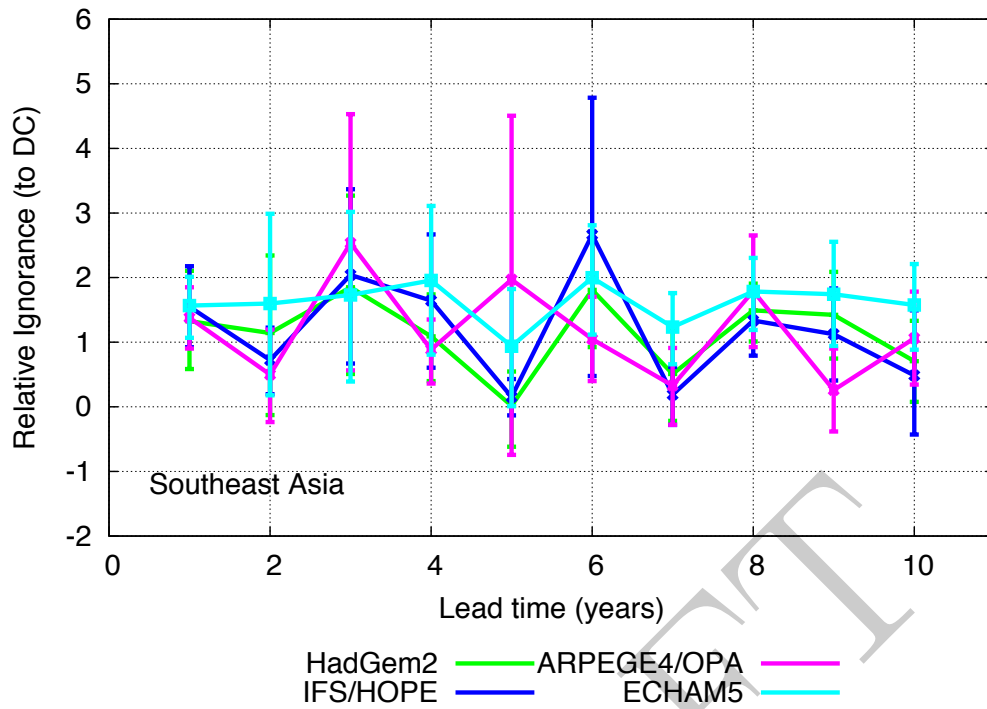


FIG. 21. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southeast Asia. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

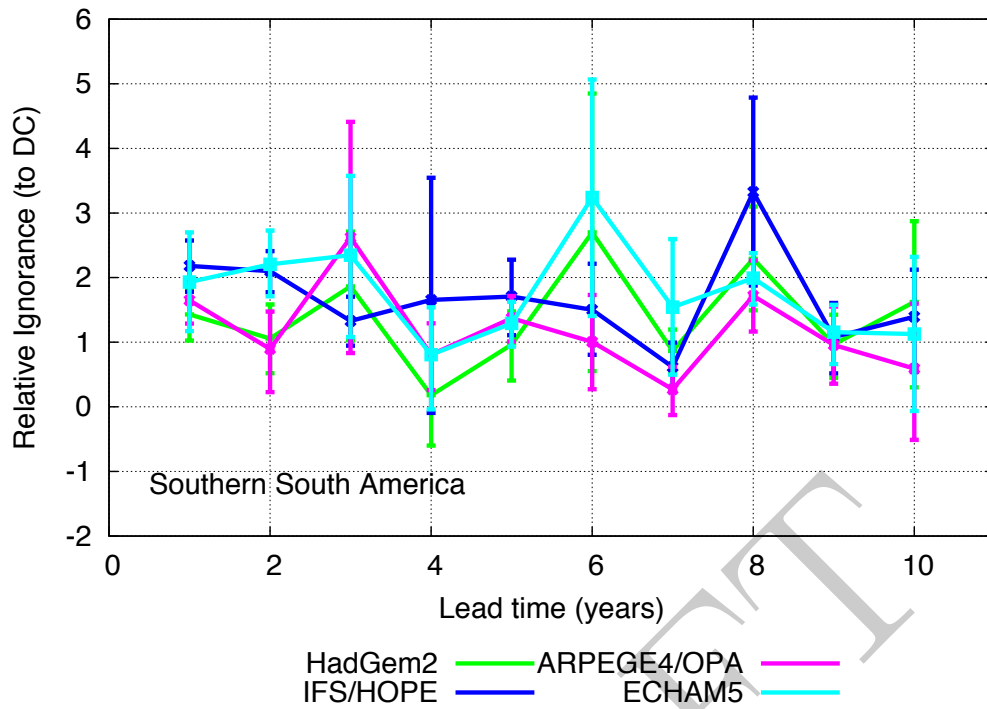


FIG. 22. Ignorance of the ENSEMBLES simulation models relative to the DC model for Southern South America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

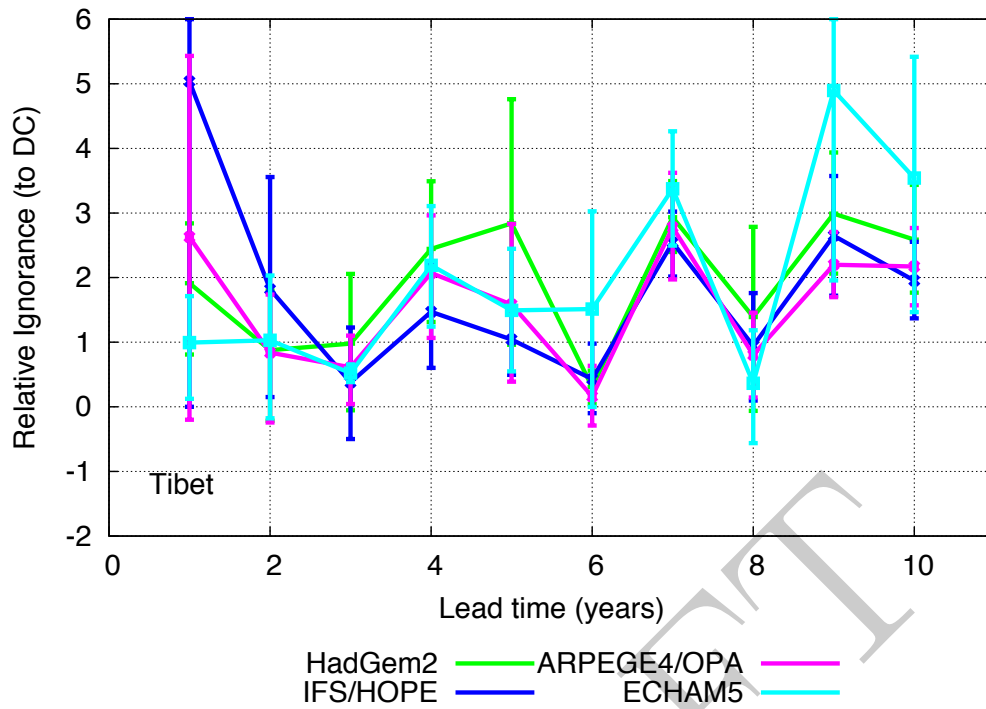


FIG. 23. Ignorance of the ENSEMBLES simulation models relative to the DC model for Tibet. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

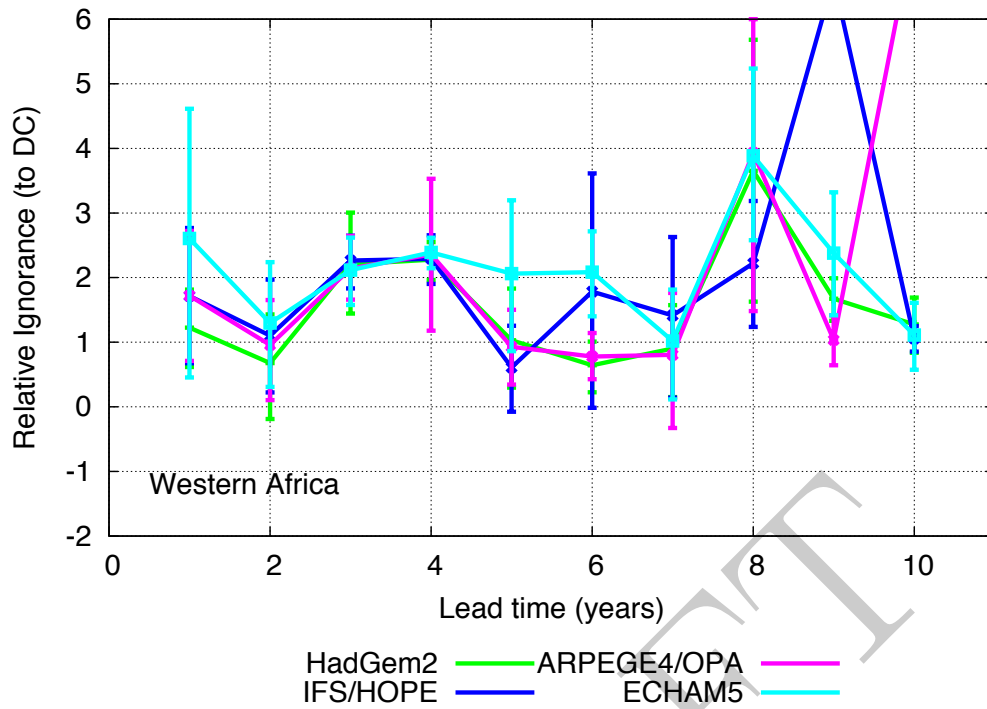


FIG. 24. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western Africa. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

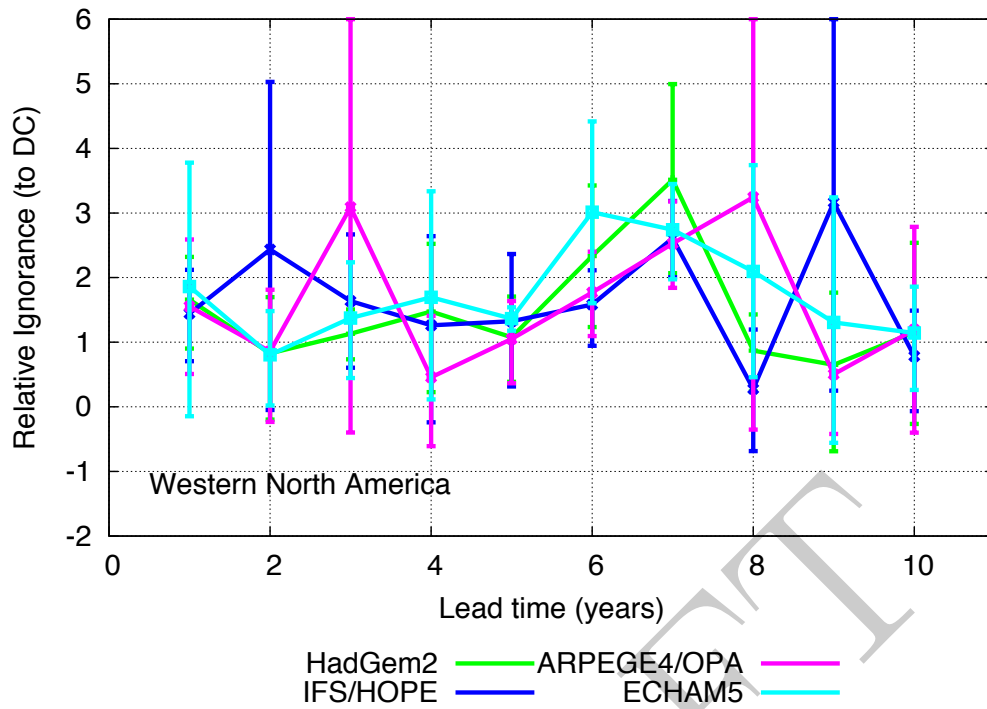


FIG. 25. Ignorance of the ENSEMBLES simulation models relative to the DC model for Western North America. Scores above zero indicate that the DC model outperforms the simulation models, placing significantly more probability on the observed outcome than the ENSEMBLES models.

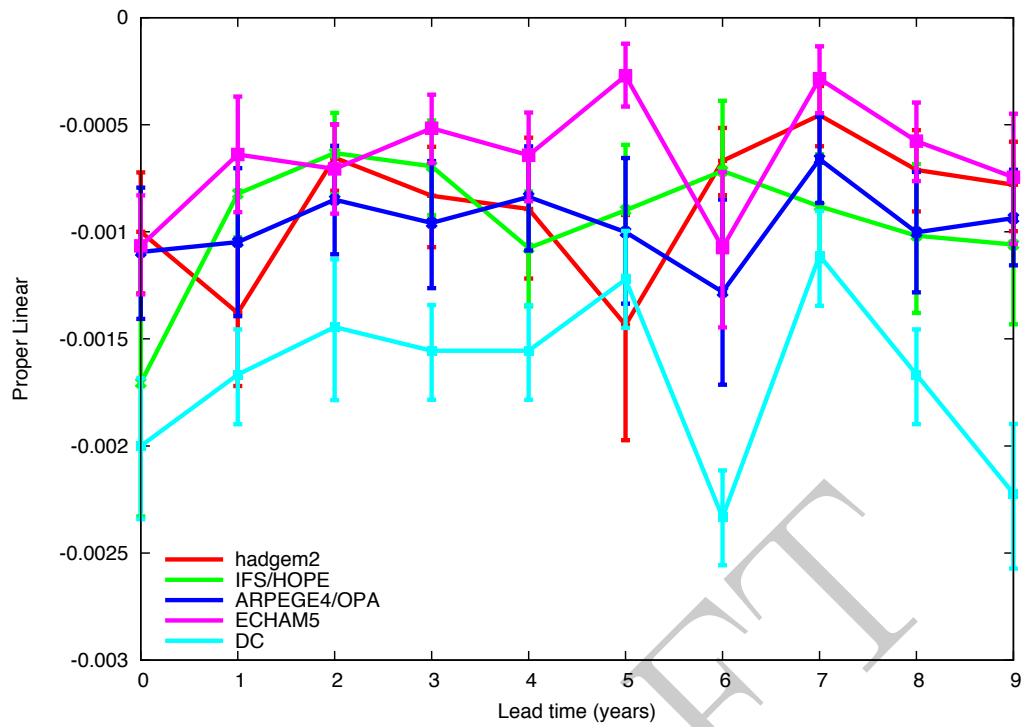


FIG. 26. Proper linear score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better forecasts. The DC model is shown to outperform the simulations models at most lead times.

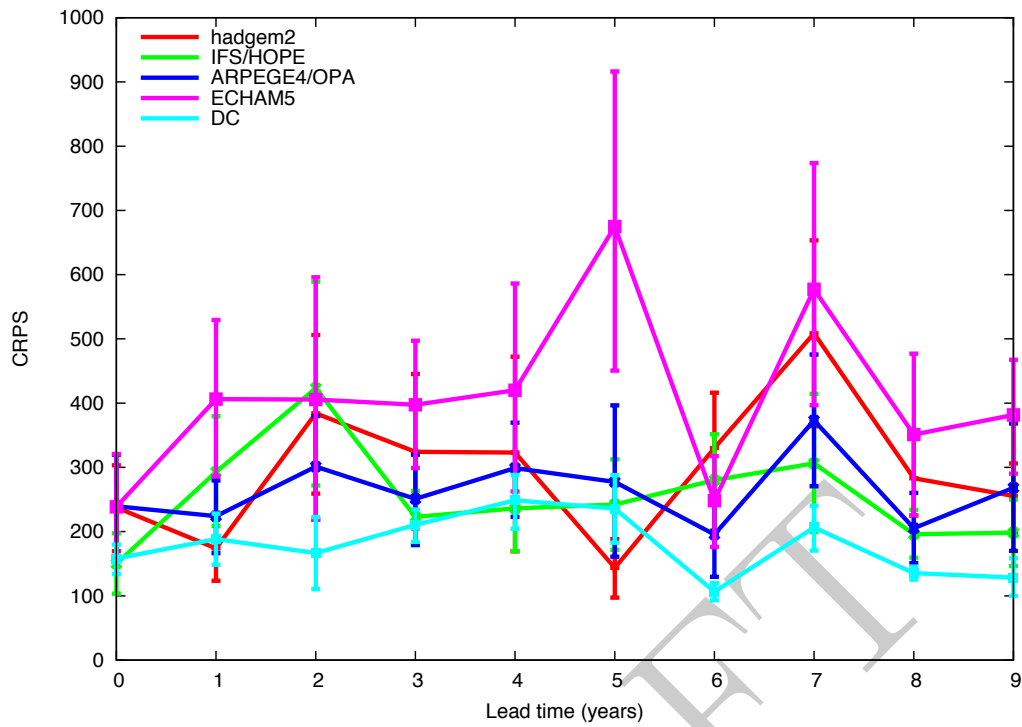


FIG. 27. CRPS score for each of the ENSEMBLES simulation models and the DC empirical model. Lower scores indicate better forecasts. The DC model is shown to outperform the simulations models at most lead times.

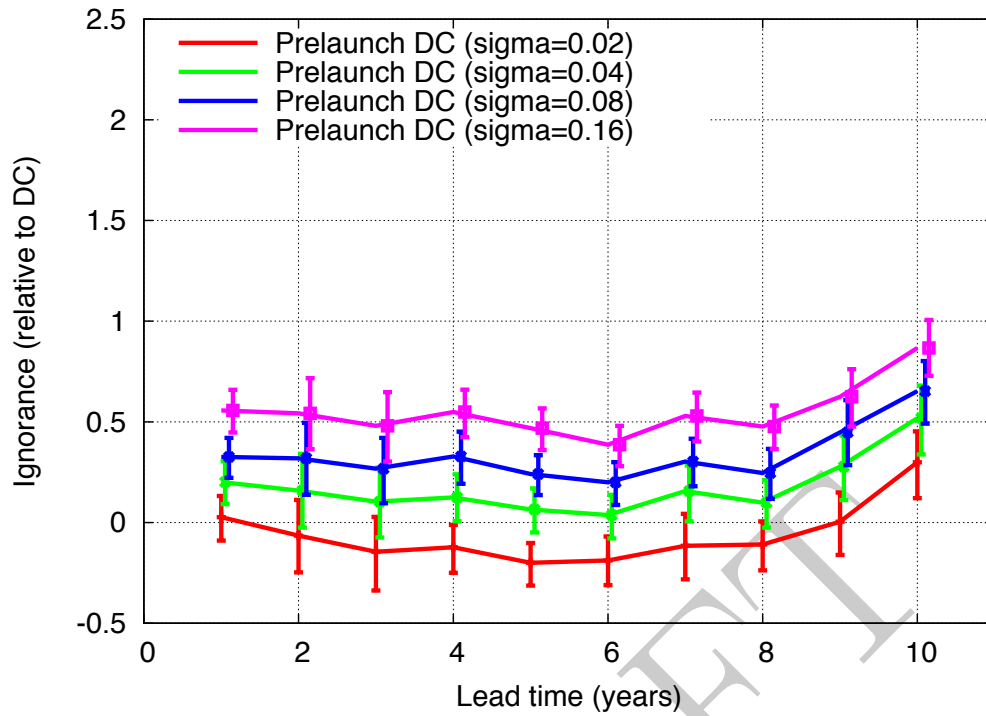


FIG. 28. Ignorance of the Prelaunch DC empirical model with kernel widths as labelled relative to the cross-validation DC model. Increasing the kernel width parameter from 0.02 to 0.16 results in a loss of skill of approximately half a bit, although for the kernel width value used in this paper (0.08)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

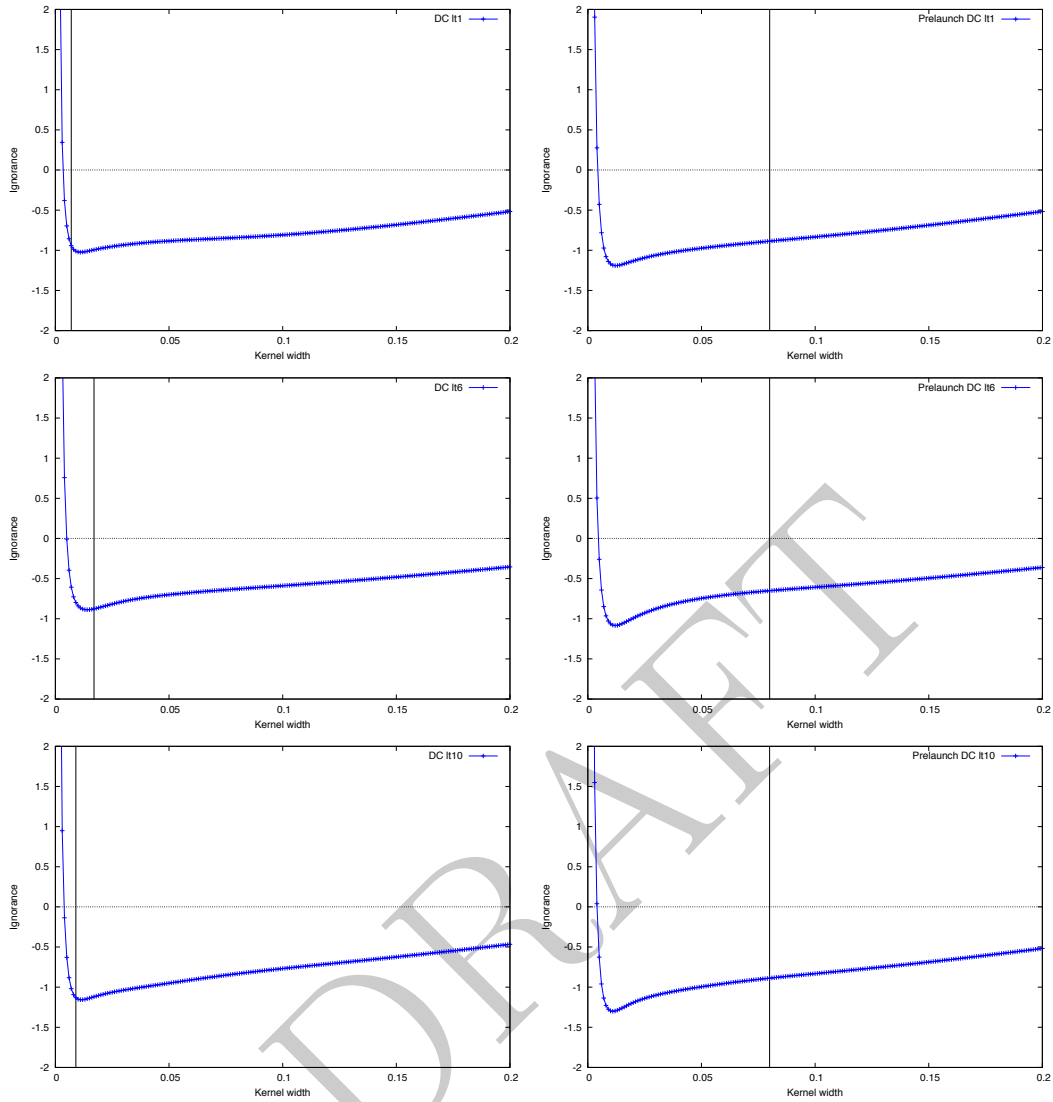


FIG. 29. Ignorance as a function of the kernel width parameter over the full set of hindcast simulations (*i.e.* with no cross-validation) for the DC (left panels) and Prelaunch DC (right panels) models at lead time one (a and b), six (c and d) and ten (e and f). The vertical bars in each case illustrate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

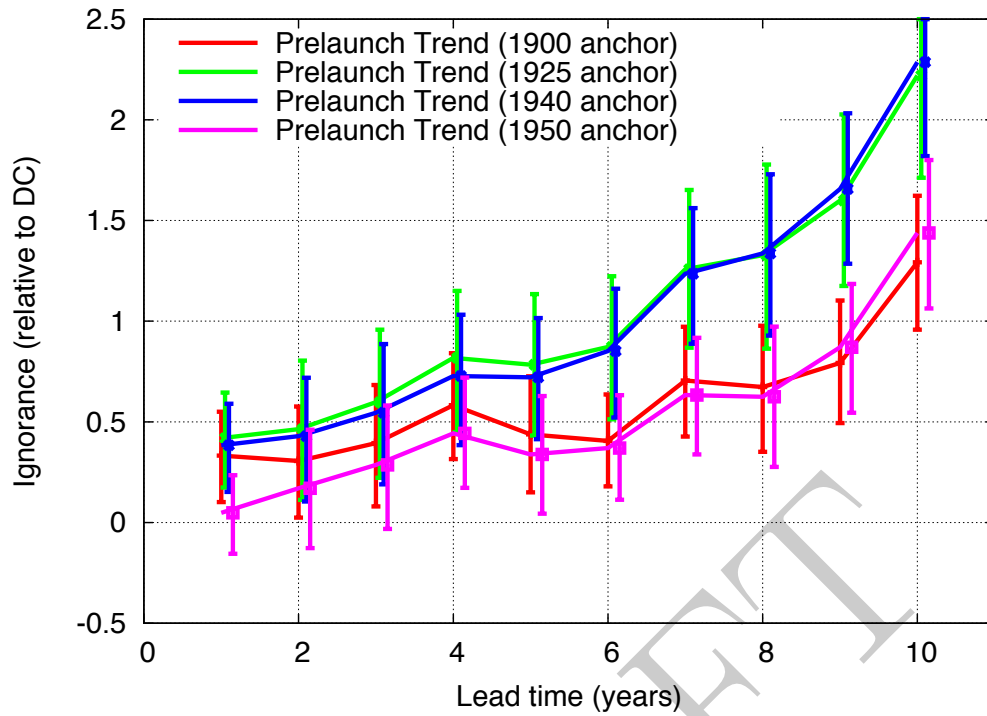


FIG. 30. Ignorance of the Prelaunch trend empirical model for different anchor times relative to the cross-validation DC model. Scores above zero indicate that DC outperforms the Prelaunch Trend model by up to half a bit at early lead times, and up to two bits (DC placing up to 4 times more probability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