An Evaluation of Decadal Probability Forecasts from State-of-the-Art Climate Models*

EMMA B. SUCKLING AND LEONARD A. SMITH

Centre for the Analysis of Time Series, London School of Economics, London, United Kingdom

(Manuscript received 26 July 2012, in final form 12 May 2013)

ABSTRACT

While state-of-the-art models of Earth's climate system have improved tremendously over the last 20 years, nontrivial structural flaws still hinder their ability to forecast the decadal dynamics of the Earth system realistically. Contrasting the skill of these models not only with each other but also with empirical models can reveal the space and time scales on which simulation models exploit their physical basis effectively and quantify their ability to add information to operational forecasts. The skill of decadal probabilistic hindcasts for annual global-mean and regional-mean temperatures from the EU Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES) project is contrasted with several empirical models. Both the ENSEMBLES models and a ''dynamic climatology'' empirical model show probabilistic skill above that of a static climatology for global-mean temperature. The dynamic climatology model, however, often outperforms the ENSEMBLES models. The fact that empirical models display skill similar to that of today's state-of-the-art simulation models suggests that empirical forecasts can improve decadal forecasts for climate services, just as in weather, medium-range, and seasonal forecasting. It is suggested that the direct comparison of simulation models with empirical models becomes a regular component of large model forecast evaluations. Doing so would clarify the extent to which state-of-the-art simulation models provide information beyond that available from simpler empirical models and clarify current limitations in using simulation forecasting for decision support. Ultimately, the skill of simulation models based on physical principles is expected to surpass that of empirical models in a changing climate; their direct comparison provides information on progress toward that goal, which is not available in model–model intercomparisons.

1. Introduction

State-of-the-art dynamical simulation models of Earth's climate system[1] are often used to make probabilistic predictions about the future climate and related phenomena with the aim of providing useful information for decision support (Anderson et al. 1999; Met Office 2011; Weigela and Bowlerb 2009; Alessandri et al. 2011; Hagedorn et al. 2005; Hagedorn and Smith 2009; Meehl et al. 2009; Doblas-Reyes et al. 2010, 2011; Solomon et al. 2007; Reifen and Toumi 2009). Evaluating the performance of such predictions from a model or set of models is crucial not only in terms of making scientific progress but also in determining how much information may be available to decision makers via climate services. It is desirable to establish a robust and transparent approach to forecast evaluation, for the purpose of examining the extent to which today's best available models are adequate over the spatial and temporal scales of interest for the task at hand. A useful reality check is provided by comparing the simulation models not only with other simulation models but also with empirical models that do not include direct physical simulation.

Decadal prediction brings several challenges for the design of ensemble experiments and their evaluation (Meehl et al. 2009; van Oldenborgh et al. 2012; Doblas-Reyes et al. 2010; Fildes and Kourentzes 2011; Doblas-Reyes et al. 2011); the analysis of decadal prediction

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

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-12-00485.s1.

Corresponding author address: Emma B. Suckling, Centre for the Analysis of Time Series, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.
E-mail: cats@lse.ac.uk

DOI: 10.1175/JCLI-D-12-00485.1

systems will form a significant focus of the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5). Decadal forecasts are of particular interest both for information on the impacts over the next 10 years, as well as from the perspective of climate model evaluation. Hindcast experiments over an archive of historical observations allow approaches from empirical forecasting to be used for model evaluation. Such approaches can aid in the evaluation of forecasts from simulation models (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012) and potentially increase the practical value of such forecasts through blending forecasts from simulation models with forecasts from empirical models that do not include direct physical simulation (Bröcker and Smith 2008).

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

A set of decadal simulations from the Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES) experiment (Hewitt and Griggs 2004; Doblas-Reyes et al. 2010), a precursor to phase 5 of the Coupled Model Intercomparison Project (CMIP5) decadal simulations (Taylor et al. 2009), is considered. The ENSEMBLES probability hindcasts are contrasted with forecasts from empirical models of the static climatology, persistence, and a ''dynamic climatology'' model developed for evaluating other dynamical systems (Smith 1997; Binter 2012). Ensemble members are transformed into probabilistic forecasts via kernel dressing (Bröcker and Smith 2008); their quality is quantified according to several proper scoring rules (Bröcker and Smith 2006). The ENSEMBLES models do not demonstrate significantly greater skill than that of an empirical dynamic climatology model either for global-mean temperature or for the land-based Giorgi region[2] temperatures (Giorgi 2002).

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

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

2. Decadal prediction systems

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

In a changing climate, even out-of-sample skill is no guarantee of future performance, because of the nonlinear nature of the response to external forcing (Smith 2002; Reifen and Toumi 2009; Solomon et al. 2007). Nevertheless, the fact that only simulation models based on the appropriate physical principles are expected to be able to generalize to new physical conditions provides no evidence that today's state-of-the-art simulation models can do so. Contrasting probability forecasts from simulation models with those from empirical models is one guide

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

to gauging the additional information derived from the physical basis of the simulation model-based forecasts. In practice, the most skillful probability forecast is often based on combining the information from both simulation models and empirical models (Van den Dool 2007; Hoeting et al. 1999; Unger et al. 2009; Bröcker and Smith 2008; Met Office 2011).

Decadal predictions aim to accurately represent both[3] the intrinsic variability and forced response to changes in the Earth system (Meehl et al. 2009). Decadal simulation models now assimilate observations of the current state of the Earth system as initial conditions in the model (Pierce et al. 2004; Troccoli and Palmer 2007). At present it is not clear whether initializing the model with observations at each forecast launch improves the skill of decadal forecasts (Pohlmann et al. 2009; Hawkins et al. 2011; Smith et al. 2007; Keenlyide et al. 2008; Smith et al. 2010; van Oldenborgh et al. 2012; Kim et al. 2012). At a more basic level, the ability to provide useful decadal predictions using simulation models is yet to be firmly established. Probabilistic hindcasts, based on simulations from stream 2 of the ENSEMBLES project [further details of which can be found in Doblas-Reyes et al. (2010) and in the appendix], do not demonstrate significantly more skill than that of simple empirical models.

Figure 1 illustrates the 2 years running mean of simulated global-mean temperature from the four simulation models in the multimodel ensemble experiment of the ENSEMBLES project over the full set of decadal hindcasts. Observations from the Hadley Centre/Climatic Research Unit, version 3 (HadCRUT3) dataset and the 40 years European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) are shown for comparison; HadCRUT3 is used as the verification dataset outcome archive for both the model evaluation and construction of the empirical model. Using ERA-40 for the verification instead of HadCRUT3 does not change the conclusions about the model skill significantly (results not presented here). Global-mean temperature is chosen for the analysis as simulation models are expected to perform better over larger spatial scales (Solomon et al. 2007). Even at the global scale, the raw simulation output is seen to differ from the observations both in terms of absolute values, as well as in dynamics. Three of the four models display a substantial model drift away from the observed global-mean temperature, with ECHAM5 being the exception. The fact that some of the models exhibit a substantial drift but not others reflects the fact that different models employ different initialization schemes (Keenlyide et al. 2005). ECHAM5 both assimilates anomalies and forecasts anomalies. Assimilating anomalies is intended to reduce model drift[4] (Pierce et al. 2004); the remaining models are initialized from observed conditions.

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

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

3. Empirical models for decadal prediction

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

[3] In reality, of course, no such distinct entities exist given the nonlinearity of the Earth system. The nature of intrinsic variability is inextricably linked to the state of the Earth system; there is no separation into a natural component and a forced component.

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

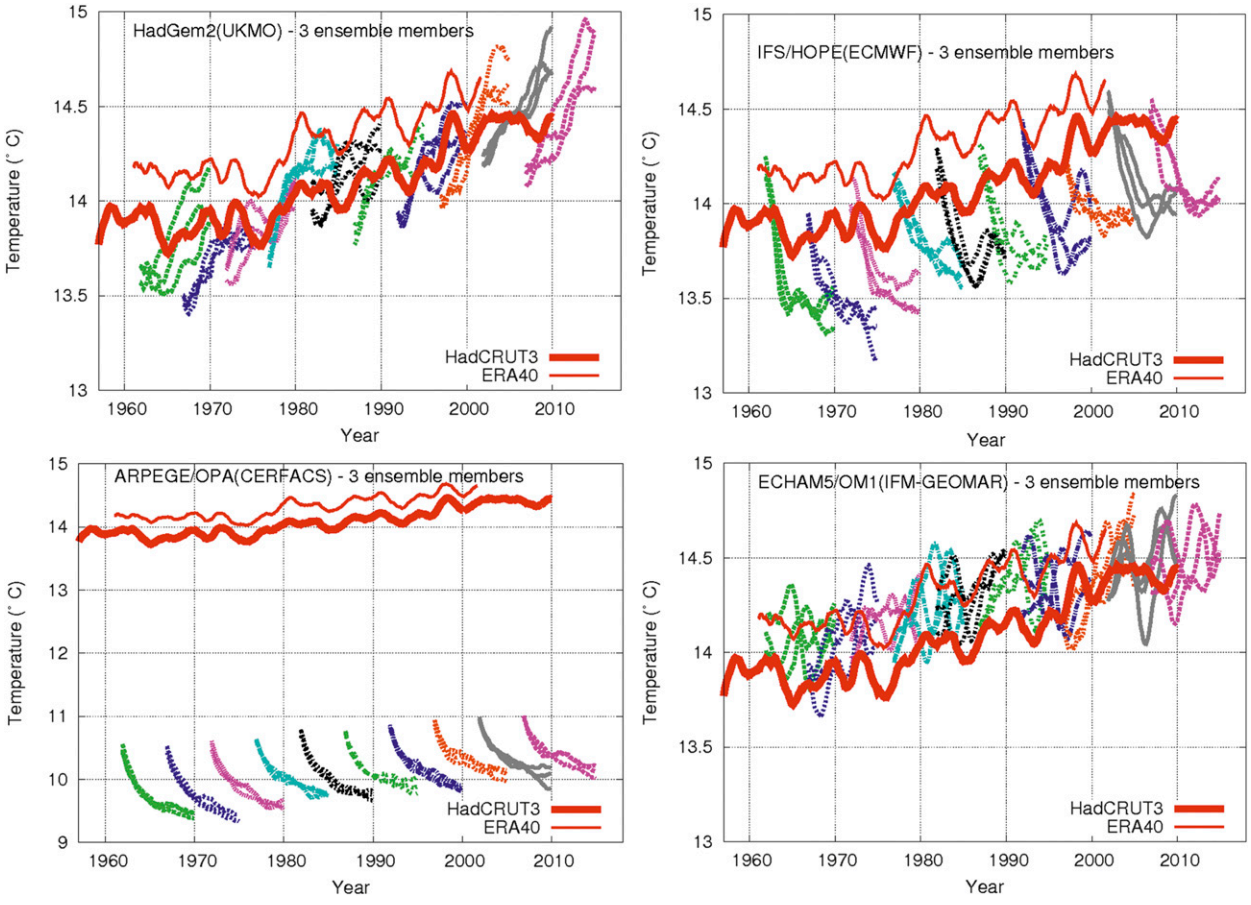

FIG. 1. Global-mean temperature (2 years running mean) for the four forecast systems: (top left) HadGEM2 [Met Office (UKMO)], (top right) IFS/HOPE (ECMWF), (bottom left) ARPEGE4/OPA [Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS)], and (bottom right) ECHAM5 [Leibniz-Institut für Meereswissenschaften (IFM-GEOMAR)] that form stream 2 of the ENSEMBLES decadal hindcast simulations (Doblas-Reyes et al. 2010). HadCRUT3 observations and ERA-40 are also shown for comparison. Note that the scale on the vertical axis for the ARPEGE4/OPA model is different than for the other three models, reflecting the larger bias in this model.

Smith 1997; Met Office 2011; Hagedorn and Smith 2009).

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

Analog forecasting uses the current state (perhaps with other recent states; Smith 1997) to define analogs within the observational record (Van den Dool 1994; Lorenz 1963; Van den Dool 2007). A distribution based on images of each analog state (the observation immediately following the analog state) then defines the ensemble forecast. Analogs may be defined in a variety of ways, including near neighbors either in observation space or in a delay reconstruction (Smith 1994, 1997). The ensemble members may be formed using the

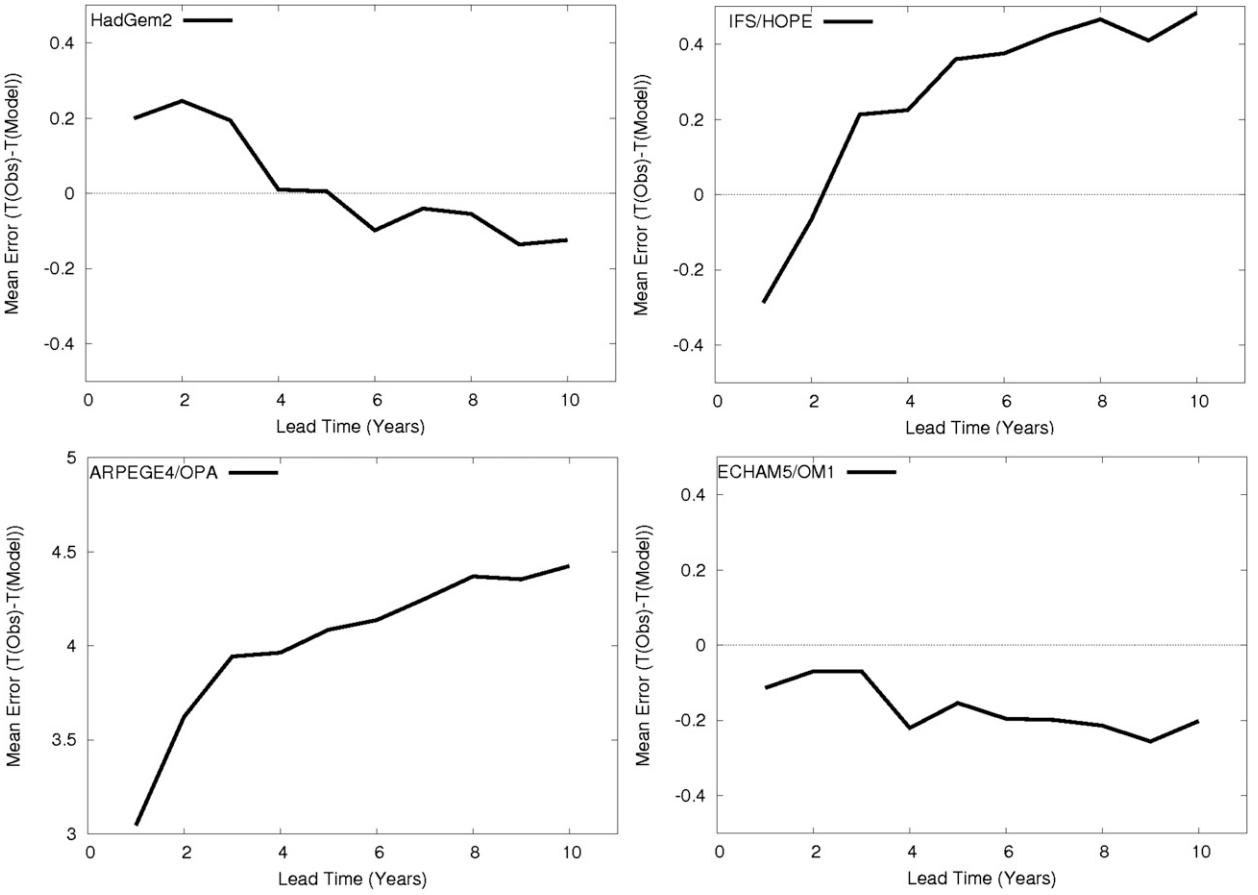

FIG. 2. Mean forecast error as a function of lead time across the set of decadal hindcasts for each of the ENSEMBLES simulation models as labeled. Note that the scale on the vertical axis for the ARPEGE4/OPA model is different than for the other models, reflecting the larger bias in this model.

complete set of available analog states (dynamic climatology; Smith 1997; Binter 2012) or by selecting from the nearest neighbors at random, with the probability of selecting a particular neighbor related to the distance (in the state space) between the prediction point and the neighbor (the random analog prediction method; Paparella et al. 1997).

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

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

where S_t is the initial condition at time t and ${}^{l}\Delta_i$ $i = 1, \ldots, N_l$ is the set of lth differences in the observational record. Figure 3 illustrates the DC model for global-mean temperature, launched at 5 years intervals, as in the ENSEMBLES hindcasts. A true out-of-sample forecast up to the year 2015, initialized to the observed global-mean temperature in 2004, is also included. Each lead time 1 forecast is based on an ensemble of 48 members. In real-time forecasting $N_l = N - l$, while for cross-validation purposes the ensembles in Fig. 3 use $N - l - 1$, omitting the Δ_j corresponding to the year being forecast. Thus at lead time 9 each forecast is based on an ensemble of 40 members. The DC approach is shown below to outperform the ENSEMBLES models when forecasting global-mean temperature.

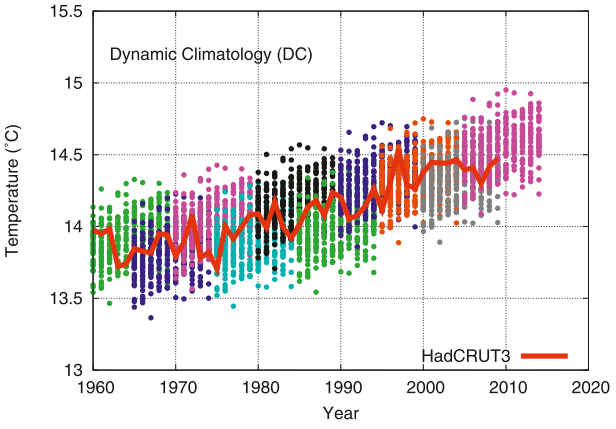

FIG. 3. The DC over the period of the ENSEMBLES hindcasts (Fig. 1). HadCRUT3 (from which the DC model is constructed) is shown for comparison.

4. Probability forecasts from ensembles

No forecast is complete without an estimate of forecast skill (Tennekes et al. 1987). Probability forecasts allow a complete description of the skill from an ensemble prediction system; they may be formed in several ways. The IPCC AR4 (Solomon et al. 2007), for example, defines a likely range subjectively, applying the "60–40" rule[5] to the mean of the CMIP3 model global-mean temperatures in 2100. Insofar as the forecast–outcome archive is larger for decadal time scales, objective statistical approaches are more easily deployed.

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

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

where x^i is the ith ensemble member, μ is the offset of the kernel mean (this offset may have a different value than the traditional bias term[6]), and σ is the kernel width. In this paper, the kernel K is taken to be a Gaussian function,

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

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

The forecasts below are evaluated using the ignorance score (Good 1952), which is defined as

$$S[p(y),Y]=-\log_2[p(Y)], \qquad (4)$$

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

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

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

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

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

[5] In chapter 10.5.4.6 of the AR4 (Solomon et al. 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" (Solomon et al. 2007, p. 810). Similar results are shown in Fig. 5 of the summary for policy makers.

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

[7] Information contamination occurs when critical information is used in a hindcast that 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 are known, principled use of cross validation can reduce its likely impact.

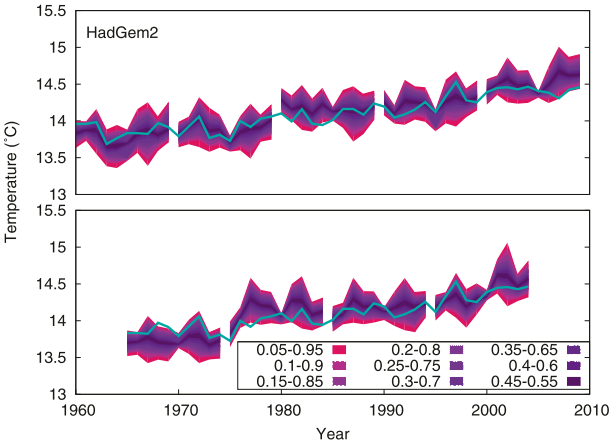

FIG. 4. Forecast distributions for HadGEM2 (UKMO) for the 5th–95th percentile. The HadCRUT3 observed temperatures are shown in blue. The forecasts are 10 years long and launched every 5 years, and so the fan charts would overlap; to avoid this they are presented in two panels: forecasts launched in 10 years intervals from (top) 1960 and (bottom) 1965.

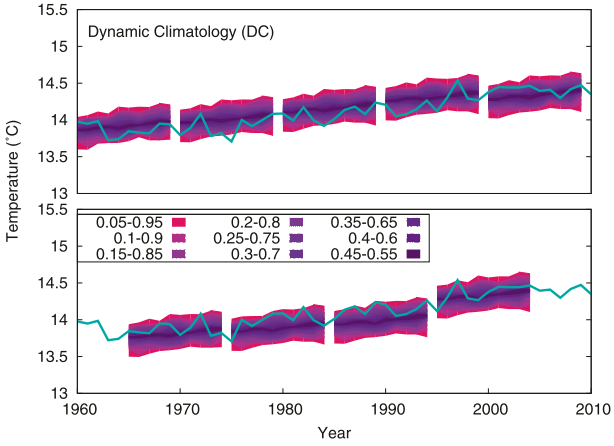

FIG. 5. As in Fig. 4, but for every fifth launch from the DC model.

The ENSEMBLES forecast–outcome archive contains at most nine forecast–outcome pairs. That is, there are only nine forecast launch dates, each with a maximum lead time of 10 years. Outside of true out-of-sample evaluation, it is difficult not to overfit the forecast and dressing parameters used to generate probability forecasts; the details of cross validation can have a large impact. Extending the typical leave-some-out fitting protocol (Hastie et al. 2001; Bröcker and Smith 2008) to include the kernel dressing procedure reduces the sample size of the forecast–outcome archive from eight to seven pairs. This ''true leave-some-out'' procedure (Smith et al. 2013, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*) will necessarily increase the sampling uncertainty, reflected through bootstrap resampling. In the case of the ENSEMBLES forecasts, adopting a true leave-some-out procedure reduces the apparent significance of the results; failing to introduce such a procedure, however, risks both information contamination and the suggestion that there is more skill than is to be expected in the simulation models. The most appropriate path cannot be determined with confidence until additional data become available.

5. Results

The skill of each of the four ENSEMBLES decadal prediction models has been evaluated relative to DC. The Hadley Centre Global Environment Model, version 2 (HadGEM2) forecast distributions are shown as fan charts in Fig. 4 as an example. These forecast distributions tend to capture the observed global-mean temperature, although the verification falls outside the 5th–95th percentile of the distribution more often than the expected 10% of the time. The distributions from the other ENSEMBLES simulation models (illustrated in the supplementary material) produce similar results.

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

Figure 6 shows the performance of all four ENSEMBLES simulation models and the DC empirical model in terms of ignorance as a function of lead time. To test whether one model is systematically better than another requires considering the relative performance directly. The ignorance of each model is computed relative to the static climatology shown in Fig. 7. True leave-some-out cross validation is applied throughout. When the relative ignorance is less than zero, the model has skill relative to the static climatology. If the bootstrap resampling intervals of a model overlap zero, the model may be less skillful than the static climatology. In fact none of the simulation models consistently outperform the DC empirical model, which has among the lowest ignorance scores. Figure 6 shows that the DC model significantly outperforms the static climatology across all lead times, on average placing approximately twice

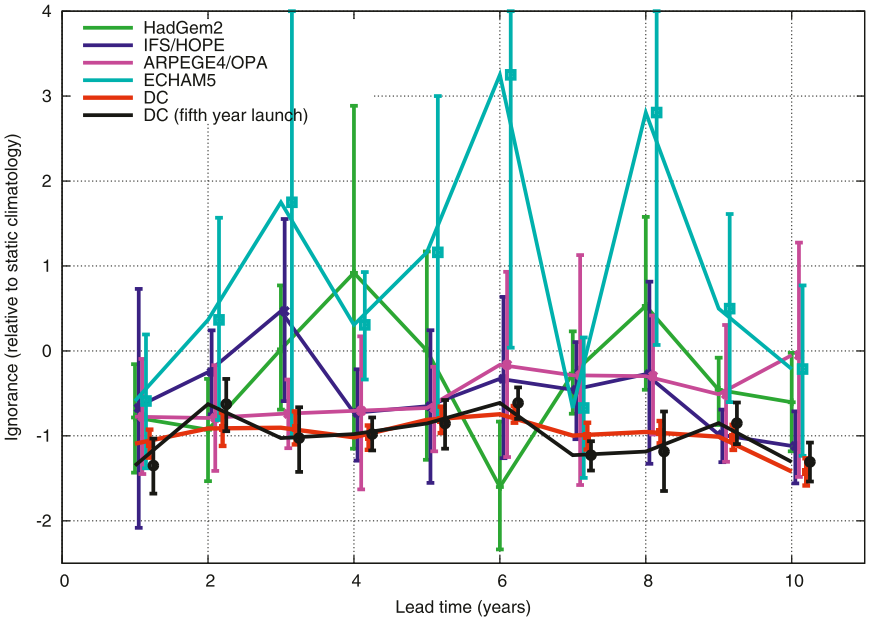

FIG. 6. Ignorance as a function of lead time for each of the four ENSEMBLES hindcast simulation models and the DC model relative to the static climatology. The bootstrap resampling intervals are illustrated at the 10th–90th percentile level. The DC model is shown to be significantly more skillful than 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.

the probability mass on the verification ($S_{\text{rel}} \approx -1.0$). The DC model using only launch dates every fifth year (to introduce a sampling uncertainty comparable to those of the ENSEMBLES model forecasts) shows a similar result but with slightly larger bootstrap resampling intervals as expected. For each of the ENSEMBLES models variations in skill between forecasts (for a given lead time) prevent the establishment of significant skill relative to the static climatology, despite the fact that both the Integrated Forecast System

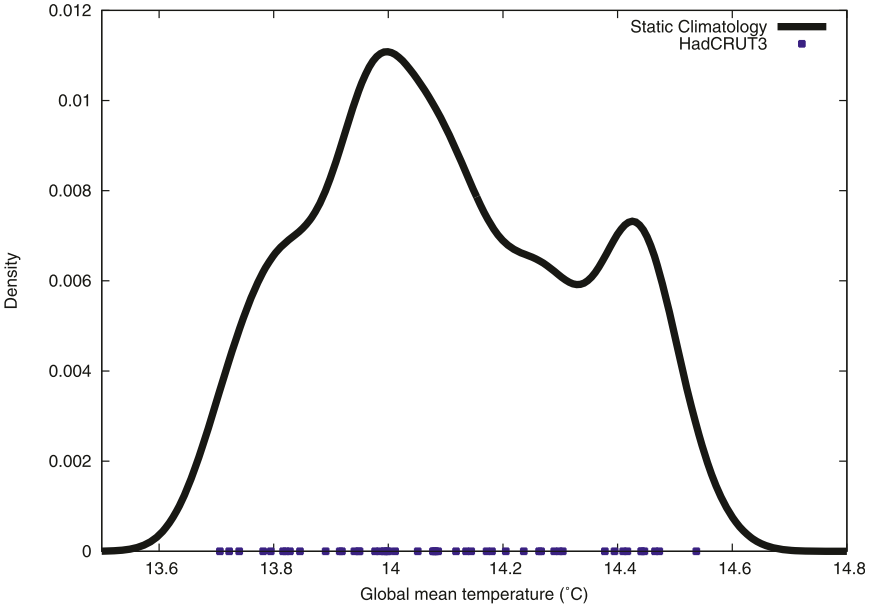

FIG. 7. Probability density for the static climatology used in the paper with observations over the period 1960–2010 (from HadCRUT3) illustrated as dots on the x axis for reference.

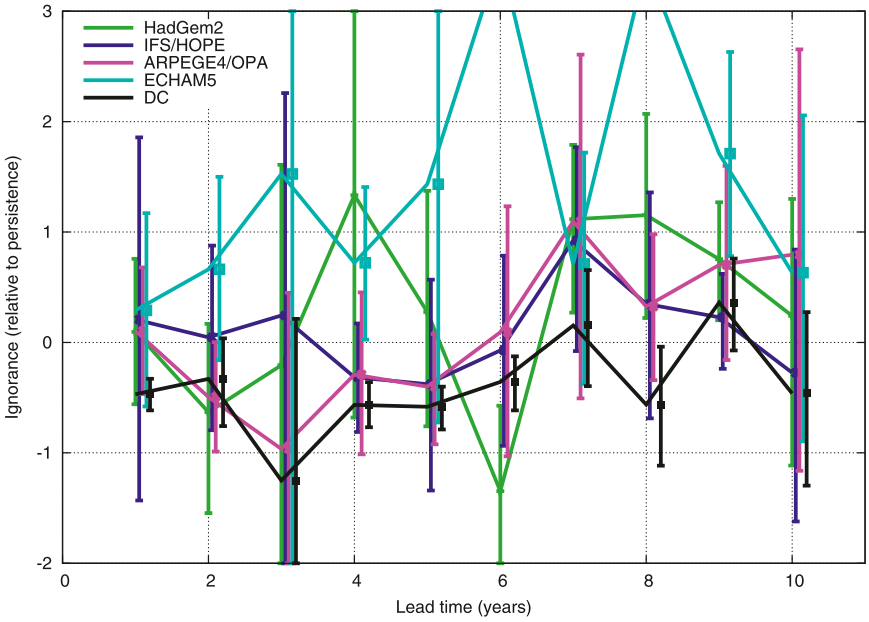

FIG. 8. Ignorance of the ENSEMBLES models and DC relative to persistence forecasts as a function of lead time. The DC model has negative relative ignorance scores up to 6 years ahead, indicating it is significantly more skillful than persistence forecasts at early lead times. The ENSEMBLES models tend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

(IFS)/Hamburg Ocean Primitive Equation Model (HOPE) and Action de Recherche Petite Echelle Grande Echelle (ARPEGE4)/Océan Parallélisé (OPA) models consistently produce relative ignorance scores below zero at most lead times. The HadGEM2 and ARPEGE4/OPA models, however, indicate that significant skill relative to static climatology can be established for early lead times. It is no surprise that the DC model performs better than the static climatology, since an increase in skill is almost certain to come from initializing each forecast to the observed temperature value at the forecast launch.

Figure 8 shows the performance of each of the models relative to forecasts of persistence. Once again the DC model consistently shows relative ignorance scores below zero across most lead times, while the ARPEGE4/OPA model scores below zero for early lead times (up to a lead time of 5 years), suggesting that forecasts from these models are more skillful than a persistence forecast over this range. In both cases the resampling bars cross the zero relative skill axis, clouding the significance of the result.

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

The ECHAM5 model generally has the least skill out of the ENSEMBLES models, particularly for global-mean temperature, with DC outperforming this model by several bits at lead times of up to 10 years, although the bootstrap resampling intervals often overlap the zero line and also overlap with the intervals from the other simulation models in Fig. 9. At global-mean temperature scales the ARPEGE4/OPA model tends to perform better than the other ENSEMBLES models, perhaps surprisingly, since the raw simulation hindcasts from ARPEGE4/OPA contain a particularly large (but consistent) model drift relative to the other simulation models. Models requiring empirical drift corrections are less likely to produce realistic forecasts in a changing climate than they are in the current climate. Over the smaller spatial scales considered (the Giorgi regions) the ARPEGE4/OPA model no longer outperforms the other simulation models; no one ENSEMBLES model

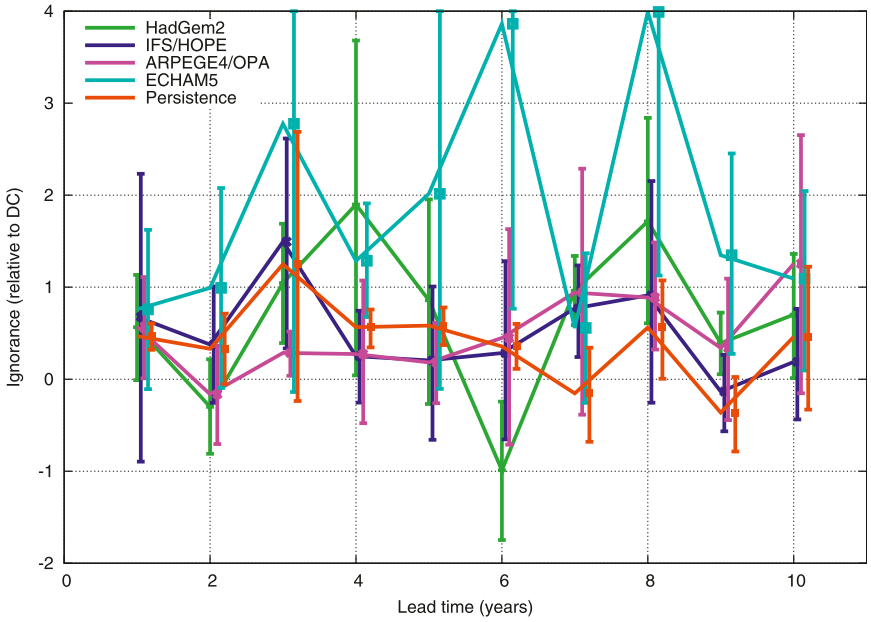

FIG. 9. Ignorance of the ENSEMBLES models relative to DC as a function of lead time. The bootstrap resampling intervals are illustrated at the 10th–90th percent level. Note that the simulation models tend to have positive scores (less skill) than the DC model at every lead time.

emerges as significantly better than any other (see supplementary material).

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

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

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

Two prelaunch empirical models were considered. The first is simply a direct climatology model where the observation archive is restricted to values prior to each launch date. The results are similar and in fact sometimes slightly better than the standard DC model.

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

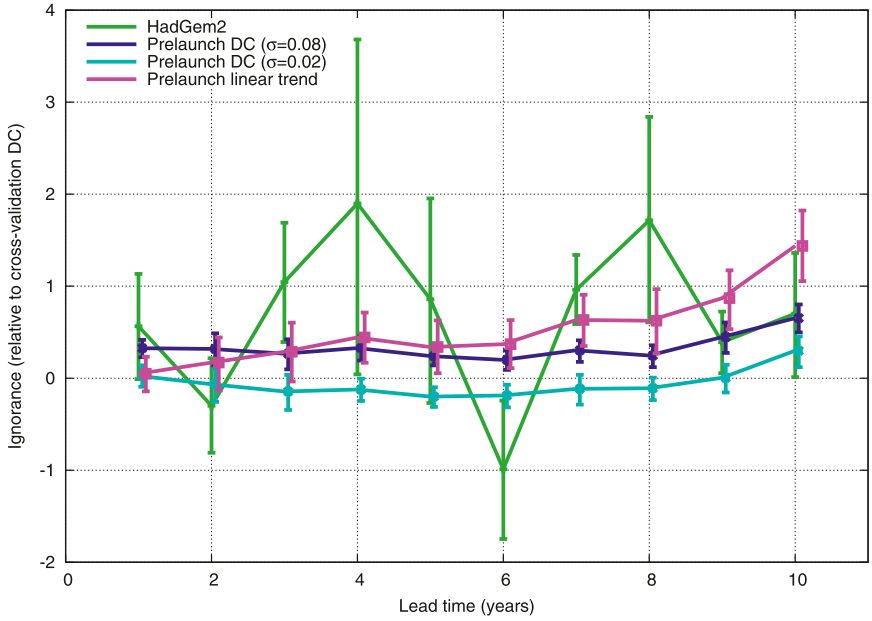

FIG. 10. Ignorance of the prelaunch DC and prelaunch trend models relative to the standard DC model as a function of lead time. The HadGEM2 model from ENSEMBLES is also shown. It is shown that the prelaunch DC model is not significantly less skillful than the standard DC model and is robust to variations 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 10th–90th percentile level.

Figure 10 shows the skill of the prelaunch DC model with a kernel width of ($\sigma = 0.08$ and 0.02) relative to the standard DC model, constructed under cross validation; performance is robust to decreasing this width by more than an order of magnitude. A prelaunch trend model was also constructed to determine if the observed skill was due to a linear trend. The prelaunch trend model simply extends the linear fit to the observations from a fixed start date (e.g., 1950) to the launch date and then uses the standard deviation of the residuals as the kernel width. The prelaunch DC is more skillful than the prelaunch trend model, as shown in Fig. 10. This result is robust to changing the start date back toward 1900 (see supplementary material). It is important to stress that this trend model is not being advocated as a candidate empirical model but only to address the specific question of whether the skill of the DC model comes only from the observed trend in global-mean temperature. Much more effective methods for estimating statistical time series models are available in this context (see, e.g., Fildes and Kourentzes 2011).

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

further into the past or where this is not possible, including intermediate launch dates to increase the sample size for the purpose of fitting the kernel dressing parameters.

6. Conclusions

The quality of decadal probability forecasts from the ENSEMBLES simulation models has been compared with that of reference forecasts from several empirical models. In general, the stream 2 ENSEMBLES simulation models demonstrate less skill than the empirical DC model across the range of lead times from 1 to 10 years. The result holds for a variety of proper scoring rules including ignorance (Good 1952), the proper linear score (PL) (Jolliffe and Stephenson 2003), and the continuous ranked probability score (CRPS) (Bröcker and Smith 2006). A similar result holds on smaller spatial scales for the Giorgi regions (see supplementary material). These new results for probability forecasts are consistent with evaluations of root-mean-square errors of decadal simulation models with other reference point forecasts (Fildes and Kourentzes 2011; van Oldenborgh et al. 2012; Weisheimer et al. 2009). The DC probability forecasts often place up to 4 bits more information (or 2^4 times more probability mass) on the observed outcome than the ENSEMBLES simulation models.

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

The evaluation and comparison of decadal forecasts will always be hindered by the relatively small samples involved when contrasted with the case of weather forecasts; the decadal forecast–outcome archive currently considered is only half a century in duration. Advances both in modeling and in observation, as well as changes in Earth's climate, are likely to mean the relevant forecast–outcome archive will remain small. One improvement that could be made to clarify the skill of the simulation models is to improve the experimental design of hindcasts: in particular, to increase the ensemble size used. For the ENSEMBLES models, each simulation ensemble consisted of only three members launched at 5 years intervals. Larger ensembles and more frequent forecast launch dates can ease the evaluation of skill without waiting for the forecast–outcome archive to grow larger.[9]

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

The comparison of near-term climate probability forecasts from Earth simulation models with those from dynamic climatology empirical models provides a useful benchmark as the simulation models improve in the future. The blending (Bröcker and Smith 2008) of simulation models and empirical models is likely to provide more skillful probability forecasts in climate services, for both policy and adaptation decisions. In addition, clear communication of the (limited) expectations for skillful decadal forecasts can avoid casting doubt on well-founded physical understanding of the radiative response to increasing carbon dioxide concentration in Earth's atmosphere. Finally, these comparisons cast a sharp light on distinguishing whether current limitations in estimating the skill of a model arise from external factors like the size of the forecast–outcome archive or from the experimental design. Such insights are a valuable product of ENSEMBLES and will contribute to the experimental design of future ensemble decadal prediction systems.

Acknowledgments. This research was funded as part of the NERC EQUIP project (NE/H003479/1); it was also supported by the EU Framework 6 ENSEMBLES project (GOCE-CT-2003-505539-ENSEMBLES) and by both by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy,

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

funded by the Economic and Social Research Council and Munich Re. L.A.S. gratefully acknowledges support of Pembroke College, Oxford. We also acknowledge the helpful comments and insights from Roman Binter, Hailiang Du, Ana Lopez, Falk Niehörster, David Stainforth, and Erica Thompson, which helped shape this work, as well as discussions with H. van den Dool, G.-J. van Oldenborgh, A. Weisheimer, and two anonymous reviewers, which improved an earlier manuscript.

APPENDIX

The Stream 2 ENSEMBLES Decadal Hindcast Experiments

The set of decadal hindcast experiments from stream 2 of the ENSEMBLES project simulations (Doblas-Reyes et al. 2010) have a similar experimental design to the seasonal hindcast experiments discussed in Weisheimer et al. (2009). The decadal hindcasts consist of a set of initial condition ensembles, containing three ensemble members, initialized at launch, from four forecast systems—ARPEGE4/OPA (CERFACS), IFS/HOPE (ECMWF), HadGEM2 (UKMO), and ECHAM5 (IFM-GEOMAR)—to produce a multimodel ensemble. A perturbed physics ensemble containing nine ensemble members from the DePreSys forecast system (based on the HadCM3 climate model) for both initialized and unassimilated simulations also forms part of the ENSEMBLES project. The hindcasts span the period 1960–2005, with simulations from each model launched at 5 years intervals, starting in November of the launch year and run over 10 years integrations. A full initialization strategy was employed for the atmosphere and ocean using realistic estimates of their observed states (except for ECHAM5, which employed an anomaly initialization scheme), with all the main radiative forcings prescribed and perturbations of the wind stress and SST fields made to sample initial condition uncertainty of the multimodel ensemble.

REFERENCES

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

Anderson, J., H. Van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Amer. Meteor. Soc.,* **80,** 1349–1361.

Barnston, A. G., and Coauthors, 1994: Long-lead seasonal forecasts—Where do we stand? *Bull. Amer. Meteor. Soc.,* **75,** 2097–2114.

Binter, R., 2012: Applied probabilistic forecasting. Ph.D. thesis, London School of Economics and Political Science, 276 pp.

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

——, and ——, 2008: From ensemble forecasts to predictive distributions. *Tellus,* **60A,** 663–678.

Colman, A., and M. Davey, 2003: Statistical prediction of global sea-surface temperature anomalies. *Int. J. Climatol.,* **23,** 1677–1697.

Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Forecast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. ECMWF Tech. Memo. 621, 45 pp.

——, M. A. Balmaseda, A. Weisheimer, and T. N. Palmer, 2011: Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: Impact of ocean observations. *J. Geophys. Res.,* **116,** D19111, doi:10.1029/2010JD015394.

Fildes, R., and N. Kourentzes, 2011: Validation and forecasting accuracy in models of climate change. *Int. J. Forecasting,* **27,** 968–995.

Giorgi, F., 2002: Variability and trends of sub-continental scale surface climate in the twentieth century. Part I: Observations. *Climate Dyn.,* **18,** 675–691.

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **14A,** 107–114.

Hagedorn, R., and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.,* **16,** 143–155.

——, F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus,* **57A,** 219–233.

Hastie, T. J., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning.* Springer, 533 pp.

Hawkins, E., J. Robson, R. Sutton, D. Smith, and N. Keenlyside, 2011: Evaluating the potential for statistical decadal predictions of sea surface temperature with a perfect model approach. *Climate Dyn.,* **37,** 2495–2509.

Hewitt, C. D., and D. J. Griggs, 2004: Ensembles-based predictions of climate and their impacts. *Eos, Trans. Amer. Geophys. Union,* **85,** 566, doi:10.1029/2004EO520005.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–417.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 254 pp.

Keenlyide, N. S., M. Latif, M. Botzet, J. Jungclaus, and U. Schulzweida, 2005: A coupled method for initializing El Niño Southern Oscillation forecasts using sea surface temperature. *Tellus,* **57A,** 340–356.

——, ——, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature,* **453,** 84–88.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.,* **39,** L10701, doi:10.1029/2012GL051644.

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

Laepple, T., S. Jewson, and K. Coughlin, 2008: Interannual temperature predictions using the CMIP3 multi-model ensemble

mean. *Geophys. Res. Lett.,* **35,** L10701, doi:10.1029/2008GL033576.

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

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

Meehl, G. A., and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Met Office, 2011: 3-month outlook for UK contingency planning. Met Office Rep., 12 pp. [Available online at http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_user_guidance-150.pdf.]

Paparella, F., A. Provenzale, L. A. Smith, C. Taricco, and R. Vio, 1997: Local random analogue prediction of nonlinear processes. *Phys. Lett.,* **235A,** 233–240.

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

Pohlmann, H., J. H. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate,* **22,** 3926–3938.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.,* **36,** L13704, doi:10.1029/2009GL038082.

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

——, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.,* **3,** 846–849.

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

——, 1994: Local optimal prediction: Exploiting strangeness and the variation of sensitivity to initial condition. *Philos. Trans. Roy. Soc.,* **348A,** 371–381.

——, 1997: The maintenance of uncertainty. *Proc. 133rd Int. School of Physics ''Enrico Fermi'' Course,* Varenna, Italy, Società Italiana di Fisica, 177–246.

——, 2002: What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA,* **4,** 2487–2492.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. L. Miller Jr., Eds., 2007: *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, 996 pp.

Stockdale, T. N., 1997: Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.,* **125,** 809–818.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the CMIP5 experimental design. CMIP5 Rep., 33 pp. [Available online at http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.]

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

Troccoli, A., and T. N. Palmer, 2007: Ensemble decadal predictions from analysed initial conditions. *Philos. Trans. Roy. Soc.,* **365A,** 2179–2191.

Unger, D., H. Van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.,* **137,** 2365–2379.

Van den Dool, H. M., 1994: Long-range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans,* **20,** 247–270.

——, 2007: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 240 pp.

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

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

Weigela, A. P., and N. E. Bowlerb, 2009: Can multi-model combination really enhance prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **135,** 535–539.

Weisheimer, A., and Coauthors, 2009: Ensembles—A new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.,* **36,** L21711, doi:10.1029/2009GL040896.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Academic Press, 704 pp.