METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **20**: 246–255 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1409

RMetS
Royal Meteorological Society

Three recommendations for evaluating climate predictions

Thomas E. Fricker, Christopher A. T. Ferro* and David B. Stephenson
College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK

ABSTRACT: Evaluation is important for improving climate prediction systems and establishing the credibility of their predictions of the future. This paper shows how the choices that must be made about how to evaluate predictions affect the outcome and ultimately our view of the prediction system's quality. The aim of evaluation is to measure selected attributes of the predictions, but some attributes are susceptible to having their apparent performance artificially inflated by the presence of climate trends, thus rendering past performance an unreliable indicator of future performance. We describe a class of performance measures that are immune to such spurious skill. The way in which an ensemble prediction is interpreted also has strong implications for the apparent performance, so we give recommendations about how evaluation should be tailored to different interpretations. Finally, we explore the role of the timescale of the predictand in evaluation and suggest ways to describe the relationship between timescale and performance. The ideas in this paper are illustrated using decadal temperature hindcasts from the CMIP5 archive. © 2013 The Authors. *Meteorological Applications* published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

KEY WORDS decadal hindcasts; ensembles; forecasts; scores; spurious skill; verification

Received 14 September 2012; Revised 9 April 2013; Accepted 10 April 2013

1. Introduction

General circulation models (GCMs) are used to study and predict the Earth's climate, and many centres around the world are now running GCMs and building large archives of climate predictions. These include forecasts (predictions issued before the predicted quantity could be determined), hindcasts (predictions issued after the predicted quantity could be determined) and projections (predictions conditioned on specific future boundary conditions that are intended to represent plausible, but not necessarily probable, future scenarios). These predictions must be viewed critically in the light of knowledge about the capabilities and limitations of the GCMs on which they are based. This knowledge is obtained by subjecting the GCMs to a series of tests, a process we call evaluation. Many methods for climate prediction evaluation are available, but so far there has been little discussion of the merits of different methods. In this paper, we review current practice in evaluating climate predictions on seasonal, decadal and multi-decadal timescales, comment on the efficacy of the methods employed, and suggest how evaluations can be adapted to account for particular features of long-range ensemble predictions in the presence of climate change. Our ideas are motivated and illustrated with decadal climate predictions, but have some relevance to predictions on all timescales.

Evaluation is performed on two, broad levels. Component-level evaluation, in which individual components of the GCM are isolated and tested independently of the rest of the model, is commonly performed at the model development stage (e.g. Randall *et al*., 2003). This can give important information about how well certain physical processes are represented in the GCM, but ultimately the GCM must be tested at a system level, where the full model is run and the results compared to observations. The raw output of a GCM is usually subjected to some post-processing prior to system-level evaluation. We refer to the GCM and post-processing system together as a *climate prediction system*.

There are several reasons to evaluate at the system level. Administrative reasons (e.g. Mason and Weigel, 2009) include monitoring the effectiveness of resources being spent on the prediction system, for example. Another reason is to diagnose problems with the climate prediction system. Examples include detecting and quantifying systematic model biases, and attempting to identify important processes that are missing or mis-specified (e.g. Randall *et al*., 2007). This information may point towards components of the GCM that need further testing and developing, and may be used to develop the post-processing system for making empirical adjustments, such as bias correction or statistical recalibration, to the model output (e.g. Stephenson *et al*., 2005; Ho *et al*., 2012). A further reason to evaluate is to provide performance-based evidence for the credibility of predictions of future climate made by the system (Parker, 2010). A quantitative assessment of the credibility of a future prediction is made by predicting some measure of performance (such as the error) of the prediction (Otto *et al*., 2012, pers. comm.). Information about the credibility of predictions might then be used to inform weightings in a multi-model ensemble (Stephenson *et al*., 2012).

System-level evaluation is performed using hindcast experiments. A typical hindcast experiment involves running the prediction system forward from one or more initialization times. Traditionally, climate prediction systems are initialized from randomly selected climate states, but state-of-the-art decadal climate prediction systems are initialized by assimilating observations of certain components of the climate system. The aim of initialization is to allow some elements of the

* Correspondence: C. Ferro, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK. E-mail: c.a.t.ferro@ex.ac.uk

internal variability of the climate system to be predicted (Meehl *et al*., 2009) and thus to reduce the uncertainty due to natural variability (Yip *et al*., 2011). The predictions made in hindcast experiments come in a variety of formats. Most GCMs are deterministic, so a prediction system that makes a single model run will produce *point predictions* (otherwise known as deterministic predictions). Some prediction systems run the GCM several times from each initialization time with perturbations made to either the initial conditions or the model parameters in order to produce an *ensemble prediction*. Occasionally, ensemble prediction systems issue only a statistic of the ensemble, such as the mean, in the form of a point prediction. Some ensemble prediction systems go further and include within their post-processing component a method such as kernel dressing to convert the ensemble into a full *probability prediction* (e.g. Roulston and Smith, 2003). Other prediction formats are occasionally seen (for example, interval predictions), but in this paper we focus on evaluating the three described above.

After running the hindcast experiment, the next step is to evaluate the performance of the predictions by comparing them to observations. A typical evaluation involves choosing a predictand (the quantity that is predicted), choosing a lead-time (how long in advance the prediction is issued), and choosing a performance measure to summarize the correspondence between the predictions and the verifying observations of the predictand. The apparent performance of the prediction system is sensitive to all of these choices: it is usual for some predictands to be better predicted than others (for example, it is commonly found that temperatures are better predicted than precipitation); the performance of initialized prediction systems may diminish with lead-time as initial conditions are forgotten; and, since different performance measures quantify different attributes of the predictions, the apparent performance will vary according to the choice of performance measure.

Performance has many aspects, so there are several attributes we might be interested in, and evaluation should be carefully tailored towards measuring those that are considered important. We only discuss a handful of attributes here; see Jolliffe and Stephenson (2012) for many more. For point predictions, the primary attributes are accuracy and association. Accuracy is the correspondence between predictions and observations, typically quantified by some function of the magnitude of the errors such as bias or mean squared error (MSE). Association is the strength of a given relationship between predictions and observations—for example, the strength of a linear relationship, which may be measured by the correlation co-efficient. Two important attributes of ensemble and probability predictions are calibration (otherwise known as reliability) and resolution. Calibration is the correspondence between predicted probabilities and observed relative frequencies of events, while resolution concerns the variation in those observed relative frequencies stratified by the different predictions. As an example, a prediction system that always issues the climatological distribution is well calibrated, but has no resolution. Certain performance measures for probability predictions known as proper scoring rules (defined in Section 4) can be additively decomposed into terms that include a measure of reliability and a measure of resolution (Bröcker, 2009), so provide an indication of all-round performance in these attributes.

Now we review some examples of evaluation studies in the literature, highlighting the choice of predictands, lead-times and performance measures made by the authors. Randall *et al*. (2007) is devoted to the evaluation of uninitialized climate models. The focus is on assessing their ability to simulate pre-industrial climate and the post-industrial change in climate, so the lead-times are generally long (100 years or more). The annual means and standard deviations of monthly means of several atmosphere and ocean variables are considered. The evaluation is performed on ensemble means (single model ensembles and multi-model ensembles) and performance is measured by MSE and the (Pearson, product–moment) correlation between the predictions and verifying observations. Performance in simulating large scale climate phenomena (such as ENSO) is also evaluated using a variety of measures (including correlation) and graphical diagnostics. MSE and correlation of the ensemble mean are also the favoured performance measures in recent decadal prediction studies (Smith *et al*., 2007; Keenlyside *et al*., 2008; Pohlmann *et al*., 2009; Caminade and Terray, 2010; Gent *et al*., 2010; Mochizuki *et al*., 2010). In many of these studies the focus is on comparing initialized prediction systems with an uninitialized equivalent, usually in their performance in predicting multi-year averages of climate variables. The methods used in such studies have recently been consolidated in an evaluation framework proposed by Goddard *et al*. (2013), which recommends evaluating a range of averaging periods for the predictand (1- to 8-year means) using the MSE of ensemble mean point predictions. The framework also considers evaluating ensemble predictions, for which the recommendation is to convert them into probability predictions by fitting a Gaussian distribution to the ensemble, and measure performance using the continuous ranked probability score (CRPS, e.g. Hersbach, 2000). For seasonal climate predictions, the World Meteorological Organization's Commission for Basic Systems has established a Standardised Verification System for Long-Range Forecasts (SVSLRF; World Meteorological Organization, 2010) that recommends a variety of performance measures, including the MSE of ensemble means, and reliability and ROC diagrams (see Landman and Beraki, 2012 for an example of the SVSLRF in practice). The Brier and ranked probability scores are also popular in seasonal prediction (for example, Kharin and Zwiers, 2003; Weigel *et al*., 2008; Jones *et al*., 2012). Again in the context of seasonal prediction, DelSole and Shukla (2010) distinguish between skill, defined as the correspondence between predictions and their verifying observations at individual time points, and fidelity, defined as the correspondence between the prediction system's climatological distribution (i.e. the distribution of the predictand over a specific reference period) and that of the real system. They propose information theoretic measures for each of those attributes, which under certain conditions reduce to the correlation and the squared mean error of the mean of a probability for skill and fidelity respectively. Other examples of fidelity evaluation include Hudson *et al*. (2011) and Fyfe *et al*. (2011), who respectively compare the histogram and the cumulative distribution functions of ensembles pooled over a number years with those of the observations.

As we have seen, the choices made about how to evaluate vary widely between studies. Sometimes choices are influenced by the nature of the predictions under study, or the data that are available to the authors, but in many cases the choices appear to be rather arbitrary and made with little justification or testing of robustness. General evaluation frameworks have been proposed (e.g. World Meteorological Organization, 2010; Goddard *et al*., 2013) to try to reduce the level of arbitrariness, with an emphasis on simple standardized evaluation methods to allow comparison of different prediction systems. In this paper, we critique the methods commonly used in evaluation studies and propose a number of new ideas. The plan of the paper is as follows. Section 2 describes some data from the CMIP5 archive

that are used throughout the paper for illustration. Section 3 discusses the choice of performance measure. Section 4 focuses on ensemble predictions and explores how the method of evaluation should depend on how the ensemble is interpreted. In Section 5 the role of timescales in evaluation is explored. Our conclusions are summarized in Section 6.

2. Illustrative data

The ideas in this paper will be illustrated using hindcasts made by the Max Planck Institute Earth System Model (MPI-ESM-LR, Hagemann *et al*., 2012), taken from the CMIP5 archive. The model was initialized at the end of each year from 1960 to 2000 (that is, in December 1960, December 1961, ..., December 2000) and the predictions start at the beginning of each subsequent year (January 1961, January 1962, ..., January 2001) running out for the subsequent 10 years. Perturbations were made to the initial conditions at each initialization time to produce a three-member ensemble. In our examples, we consider predictions of the monthly mean global mean surface temperature anomaly, produced by averaging the globally gridded model output using cosine latitude weighting. Ideally, anomalies should be expressed relative to the model's climatological mean over a reference period that pre-dates the hindcast period (Goddard *et al*., 2013), but we do not have such data available so we express anomalies for each month relative to the mean of that month in the first year following initialization of the hindcasts. To be precise, if $X_{y,m}$ denotes the temperature for month m of year y from a hindcast initialized at the end of year $y-1$ then we define the climatology for month m to be the mean of $\{X_{y,m} : y = 1961, \ldots, 2001\}$. Verification is against the HadCRUT3 data set (Brohan *et al*., 2006), which we adjust to be anomalies relative to the observational climatological monthly means over the hindcast period 1960–2000. We have not used cross-validated climatologies (as recommended by the World Meteorological Organization, 2010) for computational simplicity and because we use these hindcasts merely to illustrate our ideas rather than to provide a definitive evaluation. Figure 1 shows the annual mean of the ensemble means of the hindcasts.

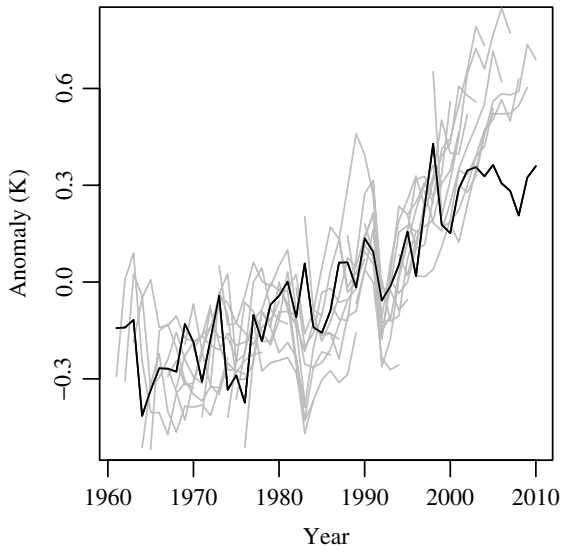

Figure 1. MPI-ESM-LR ensemble mean hindcasts of the annual mean global mean surface temperature anomaly (grey lines) and HadCRUT3 observations (black line).

3. Spurious skill and scoring rules

A performance measure is a real-valued function of a set of predictions and verifying observations. As mentioned above, performance measures should be chosen to summarize the particular attributes in which we are interested. The accuracy and association of deterministic predictions, for example, can be summarized graphically in a scatterplot and are most often measured using MSE and correlation respectively. The reliability and resolution of ensemble and probability predictions are most easily measured for predictions of binary events—exceedences of some quantile of the climatological distribution for example. These attributes may be displayed in a reliability diagram and measured by decomposing the Brier score into terms denoted REL and RES that measure reliability and resolution respectively (Murphy, 1973; Ferro and Fricker, 2012). Performance measures for continuous predictands, such as the CRPS, do have reliability/resolution decompositions (e.g. Candille and Talagrand, 2005), but these are often difficult to compute and can require larger samples of verification data than are typically available for evaluating climate prediction systems.

In a changing climate, the statistical characteristics of predictions and their verifying observations may change over the verification period. The values of some performance measures can be inflated spuriously by such changes in climate even when the attribute that the measure is designed to quantify is constant. This compromises the interpretation of the measure, provides a misleading assessment of performance and makes such measures unreliable indicators of performance in other verification periods. This phenomenon was demonstrated previously in Hamill and Juras (2006), where spurious skill appeared when weather forecasts from locations with different climates were combined. For climate predictions, spurious skill is typically induced by trends in the predictand, which means that hindcasts separated by several decades are making predictions in different climates. We show that some performance measures are susceptible to spurious skill and note that the subset of performance measures known as *scores* are immune.

Suppose that a performance measure yields a value S when we evaluate predictions over one set of verification times, T_1, and yields the same value S when we evaluate predictions (from the same prediction system) over a second set of verification times, T_2. In this situation, we may feel that combining these two data sets should only increase our confidence that the performance is indeed S. This motivates the following property for performance measures.

> *Property 1*: If the performance measured over the set of verification times T_1 is S and the performance measured over the set of verification times T_2 is S then the performance measured over the set of verification times comprising both T_1 and T_2 should also be S.

We might also want the following, stronger property to hold.

> *Property 2*: If the performance measured over the set of verification times T_1 is S_1 and the performance measured over the set of verification times T_2 is S_2 then the performance measured over the set of verification times comprising both T_1 and T_2 should lie between S_1 and S_2.

Correlation, reliability and resolution are three examples of performance measures that fail to satisfy these properties.

Table 1. Correlation and the REL and RES components of the Brier score for three time periods.

	(1962, 1981)	(1982, 2002)	(1962, 2002)
Correlation	0.43	0.75	0.76
REL	0.049	0.044	0.009
RES	0.008	0.087	0.106

Correlation and RES are positively orientated (larger values are better) and REL is negatively orientated (smaller values are better).

To demonstrate, we consider the performance of MPI-ESM-LR predictions of the annual mean global mean surface temperature anomaly in the second year of the hindcast (that is, the mean over lead-times 13–24 months), using verification years falling in three windows: $T_1 = (1962,\ 1981)$, $T_2 = (1982,\ 2002)$ and $T_3 = (1962,\ 2002)$. The correlation for the ensemble mean is shown in Table 1 and is better in the combined period, T_3, than in both T_1 and T_2. The reason for this is that the warming trend makes a positive contribution to correlation in two ways: there may be genuine skill in predicting the trend between the initialization of the hindcast and the verification time, but even unskilful hindcasts tend to correlate with their initializing observations and the trend means that they will therefore also correlate with their verifying observations. This latter contribution causes T_3 to have a better correlation than T_1 and T_2 simply because T_3 captures more of the warming trend than does T_1 or T_2. This is why we call the skill spurious.

Table 1 also shows the REL and RES components of the Brier score for predictions of the event that the temperature anomaly exceeds zero. These probability predictions are formed from the proportion of ensemble members that exceed zero. Again, the performance in T_3 is better than in both T_1 and T_2. For calibration, the prediction system tends to over-predict in T_1 and tends to under-predict in T_2. The two periods compensate for one another when they are combined, leading to a better-calibrated set of predictions. The resolution is higher in the combined period because the variability in the predictands and predictions is greater across T_3 than across either of the individual periods. Similar results (not shown) are obtained for REL and RES if the estimates are adjusted to account for the bias induced by having only three ensemble members, by using the bias-corrected estimates of Ferro and Fricker (2012).

One class of performance measures that do satisfy Properties 1 and 2 (and that are therefore, immune to spurious skill) are *scores*, which are constructed from *scoring rules*. A scoring rule is a measure that assigns a numerical value to each pair of prediction and verifying observation, and the score is the mean value of the scoring rule over all pairs. An example is the MSE for point predictions, for which the scoring rule is the squared error. However, scores are not the only performance measures that satisfy Properties 1 and 2; they are also satisfied by any monotonic function of a score, such as the square root of the MSE. However, scores do satisfy the following, yet stronger property, which weights the scores in different verification periods according to the number of data in each period.

> *Property 3*: If the performance measured over the set of N_1 verification times T_1 is S_1 and the performance measured over the set of N_2 verification times T_2 is S_2 then the performance measured over the set of verification times comprising both T_1 and T_2 should be $pS_1 + (1-p)S_2$, where $p = N_1/(N_1 + N_2)$.

That Property 3 is satisfied by scores follows immediately from the definition of scores, since the score for the combined set of verification times is $S_{12} = (N_1S_1 + N_2S_2)/(N_1 + N_2)$. Furthermore, if $S_1 = S_2 = S$ then $S_{12} = pS + (1-p)S = S$ so that Property 1 holds. Finally, S_1 and S_2 are at least as large as $\min\{S_1,\ S_2\}$ so that $S_{12} \geq p\min\{S_1,\ S_2\} + (1-p)\min\{S_1,\ S_2\} = \min\{S_1,\ S_2\}$, and S_1 and S_2 are at least as small as $\max\{S_1,\ S_2\}$ so that $S_{12} \leq p\max\{S_1,\ S_2\} + (1-p)\max\{S_1,\ S_2\} = \max\{S_1,\ S_2\}$. Therefore, Property 2 holds too.

Performance measures other than scores may still quantify relevant attributes of predictions, but we should be aware of the possibility of contamination from spurious skill and consider accounting for this in some way. One possibility described by Hamill and Juras (2006) is to evaluate the measures in sub-periods in which trends contribute little spurious skill to the performance and then pool the results appropriately. One situation in which ignoring spurious skill may be justified to some extent is when the purpose of the evaluation is to compare two or more prediction systems. For example, many of the recent decadal prediction studies (Smith *et al*., 2007; Keenlyside *et al*., 2008; Pohlmann *et al*., 2009) have focused on comparing prediction systems (for example, initialized *vs* uninitialized systems), in which case, as long as the evaluation periods are the same for both systems, the difference in measured performances will be indicative of which is superior, regardless of spurious skill. On the other hand, it may be possible for spurious skill to mask differences between the performances of prediction systems. For example, spurious skill arising from a very strong trend over the verification period could be large enough to make all correlations close to 1, in which case it may be difficult to detect differences in performance.

4. Fair performance measures

An influencing factor in deciding how to carry out an evaluation is the format in which predictions are issued. The two main formats are probability and ensemble predictions, with point predictions a special case of the latter when the ensemble size is one. In this section, we discuss how to make 'fair' evaluations of these different types of prediction. For an evaluation to be fair, we mean that the performance measure should be such that the predictor would not have wanted to change his prediction had he known that this measure were to be used. If a predictor knows the performance measure in advance then he is able to use his belief about the predictand to calculate his expectation for the values of the measure that would be obtained by issuing different predictions. A fair evaluation, therefore, uses a performance measure for which the predictor's expected value is optimized by the prediction that he did in fact issue.

In order to choose a fair performance measure, it is necessary to know (or to assume) how the issued prediction relates to the predictor's belief about the predictand. (We assume throughout that the predictor's belief corresponds to a probability distribution, which we refer to as the predictor's *belief distribution*.) This information is also important when deciding how to respond to a prediction. For example, we might decide to respond differently to a point prediction if we are told that it represents the value that the predictor considers most likely, than if we are told that it represents the value that the predictor believes will be exceeded with probability 50%. Predictions

should always be accompanied by this information, as argued by Gneiting (2011) in the case of point predictions.

For probability predictions, the most common interpretation is that the prediction coincides with the predictor's belief distribution. In other words, the predictor believes that the verifying observation will behave as if it were randomly generated from the issued probability distribution. This interpretation also seems fair if there is no other guidance available. In this case, fair evaluations are made by *proper* scoring rules, examples of which include the quadratic (Brier) score, spherical score, logarithmic score and CRPS (see Bröcker and Smith, 2007; Gneiting and Raftery, 2007; and references therein). The CRPS for a verifying observation Y and a predicted (cumulative) probability distribution function P is

$$S = \int_{-\infty}^{\infty} \{P(u) - I(Y \leq u)\}^2 \, du \tag{1}$$

where $I(Y \leq u)$ is the indicator function that equals one if $Y \leq u$ and zero otherwise. Proper scoring rules are often motivated by the fact that they encourage predictors to issue their belief distribution as the prediction (because that will optimize their expected score). Our motivation differs slightly by considering how to evaluate predictions fairly when we are in the common situation of having had no opportunity to inform the predictor in advance of him issuing his prediction which performance measure will be used for evaluation.

There are at least three conceivable interpretations of ensemble predictions. One assumes that the ensemble members represent the only values that the predictor believes may occur, and that these values are equally likely to occur. In other words, the ensemble defines the predictor's belief distribution. This is rarely the intended interpretation of an ensemble prediction, but if it were then proper scoring rules again provide fair performance measures. For example, the CRPS for a verifying observation Y and an m-member ensemble $X = \{X_1, \ldots, X_m\}$ interpreted in this way is

$$S_e = \int_{-\infty}^{\infty} \{P_m(u) - I(Y \leq u)\}^2 \, \mathrm{d}u \tag{2}$$

where $P_m(u)$ is the empirical distribution function of the ensemble, given by

$$P_m(u) = \frac{1}{m} \sum_{i=1}^{m} I(X_i \leq u). \tag{3}$$

A second interpretation assumes that the ensemble is a collection of some specified functionals (such as the mean, median or other quantiles) of the predictor's belief distribution. In other words, the predictor believes that the verifying observation will behave as if it were randomly generated from a probability distribution for which certain functionals are given by the ensemble members. A property of some functionals is *elicitability* (Gneiting, 2011). A real-valued functional F is said to be elicitable if, for any probability distribution P, there exists a performance measure S whose expected value with respect to P is uniquely optimized by the point prediction $F(P)$. Examples of elicitable functionals are the mean (elicited by the MSE performance measure) and the median (elicited by the mean absolute error). If the performance measure S elicits the functional F then S is said to be *consistent* with F and is fair. Gneiting (2011) shows that if measures are not consistent with the functional then the evaluation may favour predictions that are intuitively poor. If performance measures $S_1, \ldots, S_m$ are consistent for functionals $F_1, \ldots, F_m$ then their sum is consistent for the ensemble $\{F_1, \ldots, F_m\}$. For example, Bröcker (2012) shows that the predictor's expected value for the ensemble CRPS (Equation (2)) is optimized by choosing the ensemble, X, to be the set of quantiles of the predictor's belief distribution with probability levels $(2i-1)/(2m)$ for $i = 1, \ldots, m$.

A third interpretation of ensemble predictions assumes that the ensemble is a random sample from the predictor's belief distribution (e.g. Stephenson and Doblas-Reyes, 2000). In other words, the predictor believes that the verifying observation will behave as if it were randomly generated from the probability distribution from which the ensemble members were independently sampled. This is probably the most commonly intended interpretation of an ensemble prediction. Although ensembles are usually generated from numerical weather and climate models by systematic, rather than random, sampling of initial conditions and model parameters, the chaotic nature of the system usually means that after a short time the members are effectively a random sample. This interpretation is problematic for evaluation because no performance measure elicits random draws from the predictor's belief distribution. This is because the predictor's expected value of a performance measure is a known deterministic function of the issued prediction, and so the optimizing values for the ensemble members (if they exist) can be determined and issued as the prediction. Thus, if a prediction system is designed to optimize the expected value of any given performance measure, then it will always issue deterministic functionals of the predictor's belief distribution.

How should we evaluate an ensemble that is interpreted as a random sample? Although no performance measure will elicit a random sample, we can find measures whose expected values, assuming the predictor is forced to issue a random sample from some distribution, are optimized by choosing this distribution to be the predictor's belief distribution. The ensemble CRPS (Equation (2)) is not fair in this sense because it is optimized when the ensemble is sampled not from the predictor's belief distribution but from this distribution after it is truncated at the quantiles with levels $1/(2m)$ and $(2m-1)/(2m)$ (see the Appendix). This means that the ensemble CRPS penalizes ensemble members that are sampled from the tails of the predictor's belief distribution: the predictor is discouraged from ever predicting extreme events. In the extreme case where $m = 1$, for example, the ensemble CRPS (Equation (2)) becomes the absolute error, $|X_1 - Y|$, which elicits the median of the predictor's belief distribution. If this distribution changes little over verification times then the median will be almost constant in time. Contrast that with a sequence of random draws from the predictor's belief distribution that, while achieving a worse ensemble CRPS, will reflect the variability of the predictand. In other words, the ensemble CRPS potentially favours ensembles that we would consider bad (in the sense of having no temporal variability), over ensembles that we would consider good (in the sense of having variability that is representative of observed variability).

Although the ensemble CRPS is unfair, we can adjust it to become fair, as long as the ensemble has more than one member. The idea is to adjust the ensemble CRPS (Equation (2)) so that its expectation equals the expectation of the CRPS (Equation (1)) that would be obtained by issuing as the prediction the probability distribution from which the ensemble is generated. As the CRPS is proper, the predictor will then optimize his expected value for this adjusted CRPS by choosing the ensemble distribution to be his belief distribution. This is achieved by, for example, the adjusted version of the

ensemble CRPS proposed by Ferro *et al*. (2008),

$$S_{a,e} = S_e - \frac{1}{2m^2\,(m-1)} \sum_{i \neq j} \left| X_i - X_j \right|. \qquad (4)$$

Therefore, assuming that the predictor is restricted to issuing a random sample as his ensemble, the predictor's expected value for this adjusted ensemble CRPS is optimized by generating the ensemble from his belief distribution. Ferro (2007) and Ferro *et al*. (2008) show that the quadratic (Brier) score and (discrete) ranked probability score can also be adjusted in a similar manner. Such adjusted measures thus provide a fair way to evaluate ensembles that are assumed to be random samples.

Ferro *et al*. (2008) chose their adjustment so that the expected value of the adjusted CRPS equals the limit of the expected value of the ensemble CRPS as the ensemble size increases to infinity. Although the empirical distribution function of the ensemble converges to the ensemble distribution as the ensemble size increases, the expected value of the ensemble CRPS does not always converge to the expected value of the CRPS for the ensemble distribution. An example is when the ensemble members are exchangeable (e.g. Bröcker and Kantz, 2011) rather than independent. In such cases, the adjusted ensemble CRPS will still be unfair. In the case of independent and identically distributed ensemble members that we considered above, however, the convergence holds and the adjusted ensemble CRPS is fair. If there is only one ensemble member ($m = 1$) then the measures listed above typically cannot be adjusted to be fair (Ferro, 2007; Ferro *et al*., 2008).

Ideally, the predictor would tell us how to interpret and, therefore, use the ensemble—as its belief distribution, as a specific collection of functionals of this distribution, or as a random sample from this distribution. In practice, this information is often unavailable. For example, no such information is provided with the CMIP5 data archive. In that case, the best we can do is to choose a way to interpret the ensemble and evaluate accordingly. If predictors know that their prediction will be evaluated with a fair measure then they will be encouraged to issue the desired quantity (their belief distribution, or a specific functional of, or random sample from, their belief distribution). If unfair measures are used then we should be aware that the evaluation may favour intuitively poor predictions, and that predictors might be able to score better were they given the chance to tailor their predictions to the performance measure.

5. Timescale of the predictand

It is necessary to define the predictand (or set of predictands) to be evaluated. Multivariate predictands have been considered in a small number of studies (e.g. Smith and Hansen, 2004; Gneiting *et al*., 2008) and there is considerable potential to develop methods for evaluating predictions of high-dimensional space-time fields (e.g. Röpnack *et al*., 2013). Here, however, we follow in the steps of most authors and focus on univariate (that is, scalar) predictands. The definition of a predictand comprises three parts: a physical quantity (for example, surface temperature), a locational definition, and a temporal definition. The locational definition may be a single point, or a function over a set of points (for example, the mean over one or more grid boxes). Similarly, the temporal definition may be instantaneous or a function over a set of times (for example, the annual mean). The choices of these definitions may be restricted by the practical matter of what is actually available from the prediction and observation systems. Historic observational data sets are limited in the number of physical quantities they contain, and unless the prediction and observation systems have in-built continuous interpolators the choice of locational and temporal definitions are restricted by the spatial and temporal resolutions of those systems. Sometimes the spatial grids and time steps of the prediction and observation systems do not match, in which case some further post-processing (such as interpolation) may be required to ensure the prediction and the verifying observation are comparable. Another consideration is the quality of the available observations. Observation error increases the uncertainty in the measured performance, so it is desirable to verify against observations known to be accurate.

Beyond such practical matters, the first consideration when choosing the predictand should be whether the predictions made by the system are for a particular user. If so, then it makes sense to define the predictand to be the quantity of interest to that user. For example, Hanlon *et al*. (2012) investigate the ability of a decadal prediction system to predict heatwaves in Europe, so choose their predictands to be various extreme temperature indices that are considered relevant to the impact of heatwaves on human health. However, in many circumstances climate prediction systems are developed with many diverse users in mind (or possibly without reference to any specific user) and a more general purpose evaluation is required. This might be a first step to find strengths and weaknesses of the system, which could provide information for identifying groups of potential users, after which further testing could be performed on user-specific predictands.

In user-non-specific evaluation studies, the choice of predictand is often determined by the expected capabilities of the prediction system. For example, we have seen that in many evaluation studies authors choose as their predictand an annual or multi-year mean of the chosen physical quantity. A reason for this temporal averaging is that climate prediction implies predicting at lead-times that are longer than the lead-times found in weather prediction. Weather (that is, the instantaneous state of the atmosphere) loses predictability after a number of weeks (Lorenz, 1963), so there is no expectation that a climate prediction system will have any success in predicting weather at long lead-times. Temporal averaging is intended to filter out unpredictable short timescale variations, leaving a potentially predictable climate signal.

The choice of averaging length defines the timescale of the predictand. We expect performance to change with the timescale, and one of the goals of evaluation may be to describe these changes. First, consider predictands such as daily mean temperatures that are defined with little or no temporal averaging. Although short timescale *variations* in predictands may be unpredictable at long lead-times, there is no reason that long-range predictions of short timescale predictands cannot be well calibrated: the predictor's belief distributions must just be wider (less sharp) to encompass the unpredictable variations. Such predictions may also exhibit resolution if the predictand varies substantially over the verification period and this variation is driven by predictable long-term processes. Therefore, evaluating long-range predictions of short timescale predictands is meaningful, and may be enlightening for both developers and users of the predictions, particularly if short timescale predictands are of most relevance to the users (Smith, 2002).

It is possible that a predictor may be poor at predicting predictands defined on short timescales but good at predicting predictands defined on longer timescales, and so we should consider evaluation for longer timescales too. Let X_{li} denote ensemble member i at lead-time l from a given initialization

time, and let Y_l denote the corresponding verifying observation, where $i = 1, \ldots, m$ and $l = 1, \ldots, n$. If we compare the sets $\boldsymbol{X} = \{X_{li} : i = 1, \ldots, m \text{ and } l = 1, \ldots, n\}$ and $\boldsymbol{Y} = \{Y_l : l = 1, \ldots, n\}$ then the range, n, of the lead-times defines the timescale. Temporal averaging compares the means of the two sets: the verifying observation is defined to be $\overline{Y} = \sum_{l=1}^{n} Y_l/n$ and the prediction is the ensemble mean, $\overline{X} = \sum_{i=1}^{m} \overline{X}_i/m$, where $\overline{X}_i = \sum_{l=1}^{n} X_{li}/n$ is the time mean for ensemble member i. This prediction is often evaluated with the squared error, $\left(\overline{X} - \overline{Y}\right)^2$. An alternative prediction is the ensemble of time means, which might be evaluated using the ensemble CRPS (Equation (2)). A third option is to issue a probability prediction for $\overline{Y}$ and to evaluate that using the CRPS (Equation (1)).

Climate is more than just a time mean, however: it is the entire distribution of instantaneous weather states. Therefore, we should not limit ourselves to evaluating time means and we should consider alternative ways of comparing the sets $\boldsymbol{X}$ and $\boldsymbol{Y}$. One option is to define the verifying observation, Q_n, to be the empirical distribution function of $\boldsymbol{Y}$ and the prediction, $P_{m,n}$, to be the empirical distribution function of $\boldsymbol{X}$. This prediction could be evaluated using the scoring rule

$$S_{p,e} = \int_{-\infty}^{\infty} \left\{P_{m,n}(u) - Q_n(u)\right\}^2 \mathrm{d}u, \tag{5}$$

which we call the pooled ensemble CRPS. This reduces to the ensemble CRPS (Equation (2)) if $n = 1$. If $S_{p,e}$ is zero then the frequency distributions of ensemble members and verifying observations over the pooling interval are equal, which Gneiting *et al*. (2007) describes as marginal calibration. If probability predictions, P_l, are issued instead of ensembles for each Y_l then we could use the scoring rule

$$S_p = \int_{-\infty}^{\infty} \left\{\frac{1}{n}\sum_{l=1}^{n} P_{l(u)} - Q_n(u)\right\}^2 du, \tag{6}$$

which we call the pooled CRPS.

The ensemble scoring rules mentioned above (the squared error, ensemble CRPS and pooled ensemble CRPS) are all unfair in the sense described in Section 4. Adjusting these measures to be fair is potentially more complicated than before, however, because the necessary adjustments will depend on any temporal correlation in the verifying observations and ensemble members. We use the unadjusted scores in the remainder of this section, leaving the adjusted scores to future investigations.

A simple way of investigating the relationship between timescale and performance is to plot a temporally pooled performance measure against the width of the pooling interval. To illustrate this, we consider the MPI-ESM-LR hindcast predictions pooled over intervals of lead-times starting at 13 months and ending in the range 13–120 months. Qualitatively similar results are found for other lead-times, although performance tends to decrease as lead-time increases (not shown). Figure 2

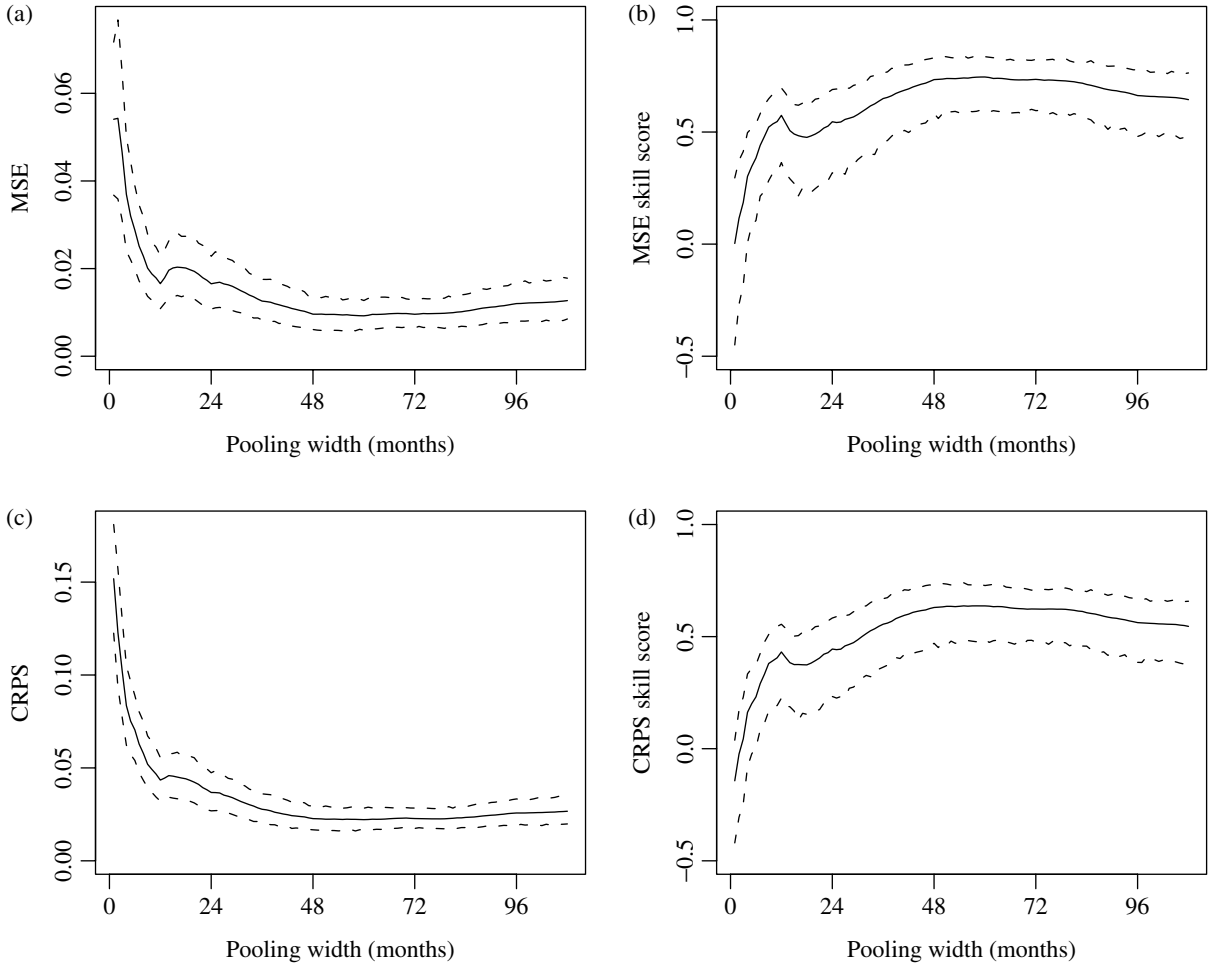

Figure 2. MSE (a), MSE skill score (b), pooled ensemble CRPS (c) and pooled ensemble CRPS skill score (d) with pointwise 90% confidence intervals (dashed lines) plotted against the width of the pooling interval.

shows the MSE and pooled ensemble CRPS (Equation (5)) plotted against the width of the pooling interval. The scores are also presented as skill scores to show the performance relative to that of an in-sample climatological reference prediction. The MSE skill score is $1-\text{MSE}/\text{MSE}_{\text{clim}}$, where MSE_{clim} is calculated for predictions equal to the climatological mean of the verifying observations over the entire hindcast period, 1961–2001. The $S_{p,e}$ skill score is $1-S_{p,e}/S_{\text{clim}}$, where S_{clim} is the pooled ensemble CRPS calculated for predictions equal to the empirical distribution function of the verifying observations over the entire hindcast period. As we are using reference predictions calculated from all of the verifying observations, these skill scores are no longer strictly scores in the sense described in Section 3, and so may be susceptible to spurious skill. Sampling uncertainty is represented in Figure 2 by pointwise, equal-tailed, percentile bootstrap, 90% confidence intervals based on non-parametric resampling of the 41 pairs of sets $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The absolute performances of the ensemble means and the ensemble distributions both tend to improve as the width of the pooling interval increases. The improvement is particularly marked as the width increases up to 12 months, and levels off by about 60 months. Performance also improves when the width is near multiples of 12 months due to greater predictability of months at the turn of the year. Some reduction in performance is noticeable for the greatest pooling widths because the ensemble members tend to over-predict substantially in the final 10 years of the verification period (Figure 1). The relative performances measured by the skill scores show similar patterns. Both skill scores are near zero for short pooling intervals, indicating that the predictions are little better than climatology for short timescale events (where natural variability dominates predictable signals), but the skill scores increase with the width of the pooling interval, indicating that performance improves faster than for the climatological predictions. The skill scores exceed zero once the temperatures are pooled over 3 months or more.

6. Summary and discussion

We have discussed some of the choices that must be made when evaluating climate prediction systems. We summarize our main points here.

- Spurious skill can arise if changes in the statistical properties of the predictions and verifying observations affect the value of a performance measure even when the attributes that the measure is designed to quantify remain unchanged. Scores (means of scoring rules) are immune to spurious skill. If measures that are susceptible to spurious skill are used then we should be aware of the potential for the measured performance to be inflated by climate trends, and for the measured performance to be an unreliable indicator of performance over other verification periods.
- Given a specific interpretation of a prediction, a performance measure is fair if a predictor would not want to have issued a different prediction had he known that his prediction would be evaluated with this measure. Fair performance measures favour predictions that perform well with respect to those attributes that we expect for predictions with the given interpretation. Proper scores are fair measures for probability predictions. Consistent measures are fair for ensemble predictions that are interpreted as a specific collection of functionals of the predictor's belief distribution. Scores such as the ensemble CRPS can be adjusted to yield fair measures for ensemble predictions that are interpreted as random samples from the predictor's belief distribution.
- Whatever the lead-time, predictions of predictands defined on all timescales can potentially perform well according to attributes such as reliability (calibration) and resolution. Evaluating predictions for predictands defined across a range of timescales (by temporal pooling or averaging, for example) is meaningful, therefore, and potentially informative for both developers and users of predictions. Predictions of short timescale predictands should not be ignored, particularly if such predictands are of most relevant to users' concerns.

The ideas that we have presented are somewhat underdeveloped and they suggest several directions for further investigation. We feel that greater understanding of different types of spurious skill, their causes and their implications for evaluation would be valuable. Many more fair performance measures for ensembles that are interpreted as different types of sample could be constructed too. We shall also be interested to see if more investigations of the performance of predictions across a wide range of timescales will help to improve the quality and utility of climate predictions.

Acknowledgements

This work was part of the EQUIP project (http://www.equip.leeds.ac.uk) funded by NERC Directed Grant NE/H003509/1. The authors thank Leon Hermanson, Doug Smith and Holger Pohlmann for useful discussion, Helen Hanlon for assistance with obtaining data, and two anonymous reviewers for comments that helped us to improve the presentation of our ideas.

Appendix

We show that if $X=\{X_1,\ \ldots,\ X_m\}$ is an m-member ensemble drawn from a distribution P then the expected value of the ensemble CRPS (Equation (2)),

$$E\left(S_e\right)=E\left[\int_{-\infty}^{\infty}\left\{P_m\left(u\right)-I\left(Y\leq u\right)\right\}^2\mathrm{d}u\right], \tag{A1}$$

is optimized when P is a truncated version of Q, the distribution of Y.

It follows from the results of Ferro *et al*. (2008) that if X is a random sample from P then the bias of the ensemble CRPS is

$$E\left(S_e\right)-S=\frac{1}{2m}E\left(\left|X_1-X_2\right|\right), \tag{A2}$$

where S is the CRPS (Equation (1)) and the expectations are taken with respect to X. Taking expectations with respect to Y, we have

$$E\left(S\right)=\int_{-\infty}^{\infty}\left\{P\left(u\right)^2-2P\left(u\right)Q\left(u\right)+Q\left(u\right)\right\}\mathrm{d}u \tag{A3}$$

and, since the ensemble CRPS reduces to the absolute error when the ensemble size is one (Gneiting and Raftery, 2007),

$$\begin{aligned}E\left(\left|X_1-X_2\right|\right)&=E\left[\int_{-\infty}^{\infty}\left\{I\left(X_1\leq u\right)-I\left(X_2\leq u\right)\right\}^2\mathrm{d}u\right]\\&=2\int_{-\infty}^{\infty}P\left(u\right)\left\{1-P\left(u\right)\right\}du.\end{aligned} \tag{A4}$$

Therefore,

$$\begin{aligned} E(S_e) &= E(S) + \frac{1}{2m} E(|X_1 - X_2|) \\ &= \int_{-\infty}^{\infty} \left[\{P(u) - Q(u)\}^2 + Q(u)\{1 - Q(u)\} + \frac{1}{m} P(u)\{1 - P(u)\} \right] du \\ &= \int_{-\infty}^{\infty} \left\{ \frac{m-1}{m} P(u)^2 + \frac{1}{m} P(u) - 2P(u)Q(u) + Q(u) \right\} du \\ &= \int_{-\infty}^{\infty} \left[\frac{m-1}{m} \left\{ P(u) - \frac{2Q(u)m - 1}{2(m-1)} \right\}^2 - \frac{\{2Q(u)m - 1\}^2}{4m(m-1)} + P(u) \right] du \\ &= \int_{-\infty}^{\infty} \left[\frac{m-1}{m} \left\{ P(u) - \frac{2Q(u)m - 1}{2(m-1)} \right\}^2 + \frac{4m^2 Q(u)\{1 - Q(u)\} - 1}{4m(m-1)} \right] du. \end{aligned} \tag{A5}$$

This expected score is optimized by minimizing the first term in the integrand, subject to the constraint that $P(u)$ is a distribution function. This is achieved by setting

$$P(u) = \begin{cases} 0 & \text{if } Q(u) < \frac{1}{2m}, \\ \frac{2Q(u)m-1}{2(m-1)} & \text{if } \frac{1}{2m} \le Q(u) < \frac{2m-1}{2m}, \\ 1 & \text{if } Q(u) \ge \frac{2m-1}{2m}. \end{cases} \tag{A6}$$

That is, the expected score is optimized when the distribution from which the ensemble is drawn is Q truncated at the quantiles with levels $1/(2m)$ and $(2m-1)/(2m)$. These are the minimum and maximum of the set of quantiles elicited by the CRPS (Bröcker, 2012).

References

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Q. J. R. Meteorol. Soc.* **135**: 1512–1519.

Bröcker J. 2012. Evaluating raw ensembles with the continuous ranked probability score. *Q. J. R. Meteorol. Soc.* **138**: 1611–1617.

Bröcker J, Kantz H. 2011. The concept of exchangeability in ensemble forecasting. *Nonlinear Processes Geophys.* **18**: 1–5.

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: the importance of being proper. *Weather Forecast.* **22**: 382–388.

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res.* **111**(D12): D12106.

Caminade C, Terray L. 2010. Twentieth century Sahel rainfall variability as simulated by the ARPEGE AGCM, and future changes. *Clim. Dyn.* **35**: 75–94.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131**: 2131–2150.

DelSole T, Shukla J. 2010. Model fidelity versus skill in seasonal forecasting. *J. Clim.* **23**: 4794–4806.

Ferro CAT. 2007. Comparing probabilistic forecasting systems with the Brier score. *Weather Forecast.* **22**: 1076–1088.

Ferro CAT, Fricker TE. 2012. A bias-corrected decomposition of the Brier score. *Q. J. R. Meteorol. Soc.* **138**: 1954–1960.

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.* **15**: 19–24.

Fyfe JC, Merryfield WJ, Kharin V, Boer GJ, Lee WS, von Salzen K. 2011. Skillful predictions of decadal trends in global mean surface temperature. *Geophys. Res. Lett.* **38**(22): L22801.

Gent PR, Yeager SG, Neale RB, Levis S, Bailey DA. 2010. Improvements in a half degree atmosphere/land version of the CCSM. *Clim. Dyn.* **34**: 819–833.

Gneiting T. 2011. Making and evaluating point forecasts. *J. Am. Stat. Assoc.* **106**: 746–762.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. B* **69**: 243–268.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**: 359–378.

Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17**: 211–235.

Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Deser C, Mason SJ, Kirtman BP, Msadek R, Sutton R, Hawkins E, Fricker T, Kharin S, Merryfield W, Hegerl G, Ferro CAT, Stephenson DB, Meehl GA, Stockdale T, Burgman R, Greene AM, Kushnir Y, Newman M, Carton J, Fukumori I, Vimont D, Delworth T. 2013. A verification framework for interannual-to-decadal predictions experiments. *Clim. Dyn.* **40**: 245–272.

Hagemann S, Loew A, Andersson A. 2012. Combined evaluation of MPI-ESM land surface water and energy fluxes. *J. Adv. Model. Earth Syst.*, DOI: 10.1029/2012MS000173 (in press).

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. R. Meteorol. Soc.* **132**: 2905–2923.

Hanlon HM, Hegerl GC, Tett SFB, Smith DM. 2012. Can a decadal forecasting system predict temperature extreme indices? *J. Clim.*, DOI: 10.1175/JCLI-D-12-00512.1 (in press).

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**: 559–570.

Ho CK, Stephenson DB, Collins M, Ferro CAT, Brown SJ. 2012. Calibration strategies: a source of additional uncertainty in climate change projections. *Bull. Am. Meteorol. Soc.* **93**: 21–26.

Hudson D, Marshall AG, Alves O. 2011. Intraseasonal forecasting of the 2009 summer and winter Australian heat waves using POAMA. *Weather Forecast.* **26**: 257–279.

Jolliffe IT, Stephenson DB. 2012. Forecast Verification*:* a Practitioner's Guide in Atmospheric Science, 2nd edn. John Wiley & Sons, Ltd: Chichester, UK.

Jones C, Carvalho LMV, Liebmann B. 2012. *Forecast skill of the South American monsoon system*. *J. Clim.* **25**: 1883–1889.

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

Kharin VV, Zwiers FW. 2003. Improved seasonal probability forecasts. *J. Clim.* **16**: 1684–1701.

Landman WA, Beraki A. 2012. Multi-model forecast skill for mid-summer rainfall over southern Africa. *Int. J. Climatol.* **32**: 303–314.

Lorenz EN. 1963. Deterministic aperiodic flow. *J. Atmos. Sci.* **20**: 130–141.

Mason SJ, Weigel AP. 2009. A generic forecast verification framework for administrative purposes. *Mon. Weather Rev.* **137**: 331–349.

Meehl GA, Goddard L, Murphy J, Stouffer RJ, Boer G, Danabasoglu G, Dixon K, Giogetta MA, Greene AM, Hawkins E, Hegerl G, Karoly D, Keenlyside N, Kimoto M, Kirtman B, Navarra A, Pulwarty R, Smith D, Stammer D, Stockdale T. 2009. Decadal prediction: can it be skillful? *Bull. Am. Meteorol. Soc.* **90**: 1467–1485.

Mochizuki T, Ishii M, Kimoto M, Chikamoto Y, Watanabe M, Nozawa T, Sakamoto TT, Shiogama H, Awaji T, Sugiura N, Toyoda T, Yasunaka S, Tatebe H, Mori M. 2010. Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proc. Natl. Acad. Sci.* **107**: 1833–1837.

Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**: 595–600.

Parker WS. 2010. Predicting weather and climate: uncertainty, ensembles and probability. *Stud. Hist. Philos. Mod. Phys.* **41**: 263–272.

Pohlmann H, Jungclaus JH, Kohl A, Stammer D, Marotzke J. 2009. Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic. *J. Clim.* **22**: 3926–3938.

Randall D, Krueger S, Bretherton C, Curry J, Duynkerke P, Moncrieff M, Ryan B, Starr D, Miller M, Rossow W, Tselioudis G, Wielicki B. 2003. Confronting models with data: the GEWEX cloud systems study. *Bull. Am. Meteorol. Soc.* **84**: 455–469.

Randall DA, Wood RA, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, Stouffer RJ, Sumi A, Taylor KE. 2007. Climate models and their evaluation. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge University Press: Cambridge and New York, NY; 589–662.

Röpnack A, Hense A, Gebhardt C, Majewski D. 2013. Bayesian model verification of NWP ensemble forecasts. *Mon. Weather Rev.* **141**: 375–387.

Roulston MA, Smith LA. 2003. Combining dynamical and statistical ensembles. *Tellus* **55A**: 16–30.

Smith LA. 2002. What might we learn from climate forecasts? *Proc. Natl. Acad. Sci.* **99**: 2487–2492.

Smith DM, Cusack S, Colman AW, Folland CK, Harris G, Murphy JM. 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science* **317**: 796–799.

Smith LA, Hansen JA. 2004. Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Weather Rev.* **132**: 1522–1528.

Stephenson DB, Coelho CAS, Doblas-Reyes FJ, Balmaseda M. 2005. Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus* **57A**: 253–264.

Stephenson DB, Collins M, Rougier JC, Chandler RE. 2012. Statistical problems in the probabilistic prediction of climate change. *Environmetrics* **23**: 364–372.

Stephenson DB, Doblas-Reyes FJ. 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus* **52A**: 300–322.

Weigel AP, Liniger MA, Appenzeller C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. R. Meteorol. Soc.* **134**: 241–260.

World Meteorological Organization. 2010. *Manual on the Global Data-processing and Forecasting System*, *Vol. I*: *Global Aspects*. World Meteorological Organization: Geneva.

Yip S, Ferro CAT, Stephenson DB, Hawkins E. 2011. A simple, coherent framework for partitioning uncertainty in climate predictions. *J. Clim.* **24**: 4634–4643.