Probabilistic Verification of Monthly Temperature Forecasts

ANDREAS P. WEIGEL, DANIEL BAGGENSTOS, AND MARK A. LINIGER

Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

FRÉDÉRIC VITART

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

CHRISTOF APPENZELLER

Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

(Manuscript received 6 February 2008, in final form 16 April 2008)

ABSTRACT

Monthly forecasting bridges the gap between medium-range weather forecasting and seasonal predictions. While such forecasts in the prediction range of 1–4 weeks are vital to many applications in the context of weather and climate risk management, surprisingly little has been published on the actual monthly prediction skill of existing global circulation models. Since 2004, the European Centre for Medium-Range Weather Forecasts has operationally run a dynamical monthly forecasting system (MOFC). It is the aim of this study to provide a systematic and fully probabilistic evaluation of MOFC prediction skill for weekly averaged forecasts of surface temperature in dependence of lead time, region, and season. This requires the careful setup of an appropriate verification context, given that the verification period is short and ensemble sizes small. This study considers the annual cycle of operational temperature forecasts issued in 2006, as well as the corresponding 12 yr of reforecasts (hindcasts). The debiased ranked probability skill score (RPSS$_D$) is applied for verification. This probabilistic skill metric has the advantage of being insensitive to the intrinsic unreliability due to small ensemble sizes—an issue that is relevant in the present context since MOFC hindcasts only have five ensemble members. The formulation of the RPSS$_D$ is generalized here such that the small hindcast ensembles and the large operational forecast ensembles can be jointly considered in the verification. A bootstrap method is applied to estimate confidence intervals. The results show that (i) MOFC forecasts are generally not worse than climatology and do outperform persistence, (ii) MOFC forecasts are skillful beyond a lead time of 18 days over some ocean regions and to a small degree also over tropical South America and Africa, (iii) extratropical continental predictability essentially vanishes after 18 days of integration, and (iv) even when the average predictability is low there can nevertheless be climatic conditions under which the forecasts contain useful information. With the present model, a significant skill improvement beyond 18 days of integration can only be achieved by increasing the averaging interval. Recalibration methods are expected to be without effect since the forecasts are essentially reliable.

1. Introduction

Probabilistic ensemble forecasts have become a standard technique in numerical weather and climate forecasting. Particularly in the context of weather and climate risk management, the probability information provided by ensembles is important in that it allows us to foresee forecast uncertainties and thus the limits of predictability. Depending on the time scales considered, and ignoring uncertainties due to model errors, ensembles sample two kinds of forecast uncertainty (Schneider and Griffies 1999; Schwierz et al. 2006): In short-term and medium-range weather forecasting, they address the forecast uncertainties arising from uncertain initial conditions. In seasonal forecasting, on the other hand, ensembles sample the distribution of states that are consistent with an evolving boundary condition, such as anomalies in sea surface temperature or even greenhouse gas forcing (Liniger et al. 2007). Both medium-range weather forecasts (e.g., Molteni et al.
1996; Buizza et al. 2005; Toth et al. 1997) and seasonal climate predictions (e.g., Anderson et al. 2003; Graham et al. 2005; Barnston et al. 2003) are meanwhile well-established techniques and operationally issued by many weather and climate services. However, on the intraseasonal or monthly time scale (i.e., in between the two “extremes” of the initial value problem of medium-range weather forecasting and the boundary value problem of seasonal forecasting) there is still a pronounced gap in prediction capability (Waliser 2006).

The central problems of monthly predictions are (i) that the monthly time scale is too long for conventional weather forecasting, since the atmospheric system has typically lost its memory of the initial conditions after less than 14 days; and (ii) that it is too short for the influence of the ocean state to differ significantly from its initial state and thus to beat persistence forecasts (Vitart 2004). While there actually are processes with pronounced variability on the intraseasonal time scale, particularly the Madden–Julian oscillation (MJO; see e.g., Madden and Julian 1971; Waliser 2006; Vitart et al. 2007) and stratosphere–troposphere interactions (e.g., Baldwin et al. 2003), and while these processes do have the potential to act as sources of predictability, it is fair to say that they are not yet well reproduced by operational global circulation models (e.g., Vitart et al. 2007; (World Climate Research Programme) WCRP 2008). It is probably for these reasons that, to the authors’ knowledge, there are currently only two weather centers that operationally issue monthly forecasts on the basis of a global circulation model. These are the European Centre for Medium-Range Weather Forecasts (ECMWF; see Vitart 2004), and the Japan Meteorological Agency (JMA; see Kurihara 2002). This is in striking contrast to the wide range of application fields in weather and climate risk management that would profit from decision support on the monthly time scale, such as agriculture (e.g., Meinke and Stone 2005), insurance companies (e.g., Zeng 2000), the energy sector (e.g., Cherry et al. 2005), and the health services (e.g., McGregor et al. 2006).

So far, only very little has been published on the actual monthly prediction skill of existing global circulation models. Most notably, Vitart (2004) has provided an overview of the design and capability of the ECMWF monthly ensemble forecasting system (MOFC). Based on the evaluation of 45 forecast samples, covering a span of 2 yr, he concludes that the MOFC performs better than climatology or persistence until day 18, and that even after 20 days of integration some regions are skillful. However, it is not clear how representative these results are, given that no confidence intervals are provided, and given that extended-range predictability may vary from year to year and, for example, strongly depends on the phase of El Niño (Shukla et al. 2000). Indeed, a longer period covering more verification years would be desirable. While the MOFC actually does provide 12 yr of reforecasts (or “hindcasts”), these data have so far not been considered for verification, apart from a brief estimation of hemispherically averaged skill based on the relative operating characteristics (ROC) skill metric, which is more a measure of potential predictability rather than of “true” prediction skill (Doblas-Reyes et al. 2005).

The central problem of evaluating a MOFC-type prediction system in a fully probabilistic way, considering all hindcast data available, is the small size of the hindcast ensembles (only 5 members) in comparison to the actual forecasts (51 members). This difference has severe consequences on any probabilistic skill metric that is sensitive to reliability. Consider for example a 5-member ensemble prediction system, with probability forecasts being issued by taking the fraction of ensemble members that predict a given event (e.g., “warmer than normal”): only a finite number of discrete probability forecasts can be issued that way, namely, 0%, 20%, 40%, 60%, 80%, and 100%. On the other hand, the observed frequency of this event, conditioned on the forecast probabilities, is in principle not restricted to these probabilities. Therefore, such small-sized ensemble hindcasts are inherently miscalibrated, or intrinsically unreliable (Richardson 2001; Weigel et al. 2007a,b; Ferro et al. 2008), leading to a strong underestimation of the true prediction skill.

It is therefore one of the objectives of this study to define and set up a fully probabilistic verification context that corrects for the problem of intrinsic unreliability without ignoring true (i.e., model induced) reliability deficits, and that allows to jointly consider small hindcast ensembles and large forecast ensembles in the verification. Building upon the work of Vitart (2004), we then apply this verification approach to systematically assess the skill of probabilistic MOFC near-surface temperature predictions in dependence of region, lead time, and season.

The paper is structured as follows: in section 2, the operational ECMWF monthly prediction system is described, together with the observational dataset used for verification. Section 3 provides a detailed description of the verification methodology applied. Section 4 gives a systematic overview of the prediction skill, and skill uncertainty, of MOFC temperature forecasts. A discussion on the potential for further skill improvement is given in section 5, and concluding remarks are provided in section 6.
2. Data

a. The ECMWF monthly prediction system

A full and detailed description of the MOFC can be found in Vitart (2004) and at ECMWF (2007). Here a short summary is provided. The MOFC is a coupled atmosphere–ocean global circulation model. The atmospheric component is the ECMWF atmospheric model Integrated Forecast System (IFS), which is run at Tl.159L62 resolution. This corresponds to a horizontal resolution of about 125 km and 62 levels in the vertical. The ocean component is the Hamburg Primitive Equation Model (HOPE; Wolff et al. 1997) with 29 levels in the vertical. The atmosphere and the ocean component are coupled through the Ocean–Atmosphere–Sea Ice–Soil coupler (OASIS; Terray et al. 1995).

Since 7 October 2004, the MOFC is initialized and integrated forward in time every Thursday to provide forecasts of the upcoming 32 days. It is set up as an ensemble with 51 members. Along with each real-time forecast, 5-member ensemble hindcasts are initialized at the same day of the previous 12 yr and equally integrated for 32 days. For example, the real-time 51-member forecast that was initialized on 19 January 2006 was accompanied by the integration of 5-member hindcasts that were initialized from data of 19 January 2005, 19 January 2004, . . . , 19 January 1994 (illustrated in Fig. 1).

The atmospheric component of the real-time forecasts is initialized from the ECMWF operational analysis; the hindcasts are initialized from the 40-yr ECMWF reanalysis (ERA-40; Uppala et al. 2005) if available (i.e., before August 2002) and from the ECMWF operational analysis afterward. Note that the operational analysis and the ERA-40 data do not share the same land surface representation (e.g., different soil roughness, different soil model versions, change in snow analysis, and different resolution). Most of those changes do not seem to have a significant impact on the monthly forecasts except those related to the changes in snow analysis that can generate spurious anomalies. However, the number of grid points affected by this inconsistency is small and should not affect the results of the paper. The ocean initialization stems from the ECMWF ocean data assimilation system (ODASYS), which is also used for the seasonal forecasts. However, since some of the data necessary for the ocean assimilation arrive with considerable delay, the ocean initial conditions lag the real time by about 12 days. To obtain a real-time estimate of the ocean conditions to be used for initialization, the ocean model is integrated from the respectively last analysis, thus “predicting” the actual ocean analysis.

The atmospheric perturbations for ensemble generation are produced in the same way as for the medium-range ensemble prediction system [i.e., by applying the singular vector method; Buizza and Palmer (1995); and including perturbations by targeting tropical cyclones; Puri et al. (2001)]. Additionally, the concept of stochastic physics (Palmer 2001) is applied to capture at least part of the uncertainties due to model formulation. The oceanic perturbations are generated in the same way as in the operational ECMWF seasonal forecasts (for more details see Vialard et al. 2005).

The ECMWF issues its monthly forecasts as weekly means; the predicted variables are averaged over four 7-day intervals (Monday–Sunday), which correspond to days 5–11 [week 1 (W1)], days 12–18 [week 2 (W2)], 19–25 [week 3 (W3)], and days 26–32 [week 4 (W4)]. To keep our evaluations consistent with the operational forecasts and with the verification of Vitart (2004), we apply the same averaging convention (see also Fig. 2) unless stated differently.

The forecasts considered in the present study cover the entire year 2006 (providing 52 samples forecast realizations), and the corresponding hindcasts cover the years 1994–2005, providing another 12 sets of 52 predictions each.
b. Observations

The forecasts are verified against ERA-40 data (Uppala et al. 2005) if available (i.e., for all hindcasts prior to August 2002) and against the ECMWF operational analysis otherwise. At least for ERA-40, several comparison studies have shown that near-surface temperatures are well represented by this dataset, justifying its use for the present study (Simmons et al. 2004; Kunz et al. 2007). For the verification, both forecasts and “observations” are interpolated on a grid with $1^\circ \times 1^\circ$ resolution. One of the problems in this context is that the resolution and thus the gridded orography of the operational analysis has repeatedly slightly changed during the verification period considered. To avoid inhomogeneities in the temperature time series, we therefore extrapolate all surface temperatures of the operational analysis onto the ERA-40 orography, applying a lapse rate of 0.006 K m$^{-1}$.

3. Verification context

a. The $RPSS_D$ skill metric

Our verification is based on a modified version of the widely used ranked probability skill score (RPSS; Epstein 1969; Murphy 1969, 1971). The classical RPSS is a squared measure comparing the cumulative probabilities of categorical forecast and observation vectors relative to a climatological forecast strategy. It is defined by

$$RPSS = 1 - \frac{\langle RPS \rangle}{\langle RPS_{CI} \rangle}, \text{ with}$$

$$RPS = \sum_{k=1}^{K} (Y_k - O_k)^2 \quad \text{and}$$

$$RPS_{CI} = \sum_{k=1}^{K} (P_k - O_k)^2. \quad (1)$$

The brackets $\langle \ldots \rangle$ denote the average of the RPS and RPS$_{CI}$ values over a given number of forecast–observation pairs. Here $Y_k$ is the $k$th component of a cumulative categorical forecast vector $\mathbf{Y}$, and $O_k$ is the $k$th component of the corresponding cumulative observation vector $\mathbf{O}$ the forecast is verified against. That is, $Y_k = \sum_{i=1}^{i} y_i$, with $y_i$ being the probabilistic forecast for the event to happen in category $i$, and $O_k = \sum_{i=1}^{i} o_i$, with $o_i = 1$ if the observation is in category $i$ and $o_i = 0$ if the observation falls into a category $j \neq i$. Analogously, $P_k$ is the cumulative climatological probability of the $k$th category. A more detailed description of the RPSS is provided in Wilks (2006). The RPSS is a favorable probabilistic skill score in that it is sensitive to distance (i.e., a forecast is increasingly penalized the more its cumulative probabilities differ from the actual outcome). Moreover, the RPSS is strictly proper, meaning that it cannot be optimized by hedging the probabilistic forecasts toward other values against the forecaster’s true belief. A big caveat of the RPSS is its strong negative bias for small ensemble sizes (e.g., Buizza and Palmer 1998; Richardson 2001; Kumar et al. 2001; Mason 2004). The reason for this bias is the intrinsic unreliability (Weigel et al. 2007a) of small ensembles, leading to inconsistencies in the formulation of the RPSS. However, particularly when the skill of large-sized ensemble forecasts is to be estimated on the basis of small-sized ensemble hindcasts, a “biasless” skill score is required (i.e., a skill score that is insensitive to ensemble size).

Müller et al. (2005) and Weigel et al. (2007a,b) have derived a debiased version of the RPSS, the so-called discrete ranked probability skill score (RPSS$_D$). The RPSS$_D$ lacks the RPSS’s strong dependence on ensemble size while retaining its favorable properties, in particular the strict propriety, making it the skill score of choice for the present study. If $K$ forecast categories are considered, and if these $K$ categories are equiprobable (as is chosen to be the case in this study), the RPSS$_D$ assumes a relatively simple analytical form and is given by

$$RPSS_D = 1 - \frac{\langle RPS \rangle}{\langle RPS_{CI} \rangle + D_0/M}, \quad \text{where} \quad (2)$$

$$D_0 = \frac{K^2 - 1}{6K}, \quad (3)$$

Unauthenticated | Downloaded 05/30/22 09:33 AM UTC
with $M$ being the ensemble size of the prediction system. If predictions of different ensemble sizes are to be jointly evaluated in one verification sample, this can be considered adequately by replacing the ensemble size $M$ in Eq. (2) by an appropriate effective ensemble size $M_{\text{eff}}$. If, for example, a set of predictions feeding into one verification consists of $N_f$ forecasts with ensemble size $M_f$ and $N_h$ hindcasts with ensemble size $M_h$ (in the MOFC we have $N_f = 1$, $M_f = 51$, $N_h = 12$, and $M_h = 5$), then $M_{\text{eff}}$ is given by (see the appendix for the derivation):

$$M_{\text{eff}} = M_h M_f \frac{N_h + N_f}{M_f N_h + M_h N_f}.$$  

(4)

b. The quantiles

In this study, MOFC forecasts of weekly averaged temperature are evaluated for three climatologically equiprobable categories: “colder than normal,” “normal,” and “warmer than normal”. The terciles separating these three categories are determined separately for each week of the year to adequately account for the seasonal cycle of temperatures. Otherwise, artificial “false skill” could be introduced as described by Hamill and Juras (2006). To account for systematic model errors and model drift (see section 4a), the terciles defining model climatology are not determined from the observations but from the pool of forecasts and hindcasts available. Since the model climatology directly after initialization is different from the state of the model after, say, three weeks of integration, the model terciles are calculated individually for each lead time considered.

To guarantee that no information from a given prediction is used in the verification process of that very prediction, we apply the quantile calculation in a leave-one-out cross-validation mode (e.g., Wilks 2006). This means, for each year considered the quantiles are calculated separately, using information from all years except the one under consideration.

The comparatively short hindcast period of the MOFC imposes a major problem for estimating terciles: to obtain terciles in weekly resolution, they need to be estimated on the basis of only 13 yr of data, which corresponds to 12 samples due to the cross validation applied. To enhance the statistical robustness, we therefore additionally include the week before and after the week of interest to estimate the terciles. That way, the effective sample size is increased from 12 to 36. This procedure is applied both for the observations and for the MOFC predictions for each lead time considered. This procedure results in a smoother seasonal cycle of the terciles (not shown).

c. Confidence intervals

To make inferences about the true value of a skill score, it is essential to estimate confidence intervals around the measured skill value (Joliffe 2007). This is particularly important in the context of the MOFC, where both the sample size (only 13 yr of hindcast data) and the ensemble size (only 5 members in the hindcasts) are low. As a straightforward nonparametric approach to obtain such confidence intervals, we apply a bootstrap method as described in Efron and Gong (1983) and Wilks (2006).

Assume, a skill score $S$ (in our case the RPSS$_D$), is calculated at a given grid point for a given week of the year. The skill score $S$ is then a function of $N_f$ forecast–observation pairs $d_1, d_2, \ldots, d_{N_f}$, with $N_f$ being the number of sample years feeding into the verification. In the present case, and applying the terminology of Eq. (4), $N_f = N_h + N_f = 13$. The sampling uncertainty of $S$ can be estimated by recomputing the value of $S$ $N_b$ times on the basis of $N_b$ random samples (with replacement) of sample size $N_g$ from $\{d_1, d_2, \ldots, d_{N_f}\}$. From the resulting distribution of $N_b$ $S$ values, the 90% percent confidence interval is then estimated by calculating the 5%–95% interquantile range. In the present study $N_b = 1000$.

Now assume that a skill score average, $\langle S \rangle$, is calculated (as will be done frequently in the remainder of this study). This could be a spatial average over the skill scores $S$ obtained at $N_S$ grid points (e.g., the RPSS$_D$ values at all European grid points for a given week of the year), or a temporal average over the skill scores $S$ obtained for $N_t$ forecast weeks (e.g., the RPSS$_D$ values of all summer weeks at a given grid point), or both. Thus, in the general case, the test statistic $\langle S \rangle$ is based on $N_g \times N_f \times N_S$ data pairs. However, following Candille et al. (2007), also in this case the resampling is performed only over the $N_t$ verification years rather than all $N_g \times N_f \times N_S$ data pairs, since that way the spatial and temporal correlation of forecast errors can be crudely accounted for. Otherwise, the number of independent samples would be overestimated and the confidence interval size underestimated.

4. Results

a. Model drift

We start our evaluations by examining the systematic absolute deviations of MOFC output from the verifying observations. Figure 3 shows the annual mean bias of 2-m temperature for forecasting weeks W1–W4. The model develops a substantial bias in many regions already in W1 (Fig. 3a). For parts of Greenland, South America, Southeast Asia, and Antarctica the model is
more than 1 K colder than the observations. Also southern Europe is affected by a cold bias. Large areas in continental North America and the polar oceans, on the other hand, show a warm anomaly of at least 1 K. Other areas of large differences between the model and the observations include many coastal regions, such as the entire western coast of the two Americas and southwest Africa. Off the coastlines, the oceans are generally too cold. During forecast weeks W2 until W4 the anomalies further amplify, while revealing the same spatial pattern as in W1 (Figs. 3b–d). In particular, the cold anomalies over Greenland are nearly doubled from W1 to W2, and over Antarctica a maximum bias of nearly 2 K is developed by W4. For the nonpolar oceans, an increase of the negative bias from 0.3 to 0.5 K on average can be detected after the full integration time. Generally the error growth saturates after about 20 forecast days. This result, as well as the stationarity of the spatial bias distribution, is consistent with the results of an analysis of the MOFC bias in 500-hPa geopotential height as described by Jung (2005). A more in-depth analysis (Baggenstos 2007) shows that the spatial distribution and magnitude of biases also reveals some visible seasonal variability, but their large-scale spatial structure remains similar throughout the year.

The observation of a pronounced model drift per se is not surprising, given that no “artificial” measures are applied to suppress model drift, and given that imbalances in the initial state of the model are not corrected (Vitart 2004). What is more surprising is that the bias is very pronounced already during W1. This contrasts with findings of Vitart (2004), who evaluated the MOFC bias of 500-hPa geopotential height and found that only after 10 days of model integration the model output begins to systematically deviate from the analysis. This indicates that processes not directly linked to circulation are probably contributing to the observed
systematic temperature errors (e.g., a wrong representation of surface processes or problems in the ocean–atmosphere coupling). While this topic is not further evaluated here, the existence of such strong biases emphasizes the need for reforecasts and an appropriate verification, so that systematic model errors can be diagnosed and corrected.

b. Prediction skill

We now proceed with the probabilistic verification of the MOFC temperature forecasts in dependence of lead time, region, and season, applying the verification context described in section 3. This is done for each grid point, each week of the year, and each of the four lead times W1–W4. We start by evaluating annually averaged skill.

1) Annual mean skill

Figure 4 shows maps of the RPSS$_D$ for forecast weeks W1–W4, averaged over all 52 weeks of the year. For W1 forecasts, prediction skill exceeds a value of 0.3 almost everywhere. Particularly high values on the order of 0.7 are assumed over the central and eastern equatorial Pacific [i.e., the region of the El Niño–Southern Oscillation (ENSO)]. On the other hand, comparatively low skill of only 0.1 or even below is found over the tropical Atlantic, Indian, and western Pacific Oceans. The low skill values over the tropics are likely because of two reasons. First, outside the ENSO region, interannual variability is very low. Indeed, the average tercile boundaries for a week in January over the equatorial Atlantic are at 299.50 and 229.80 K. This means, the difference between a warm and a cold week is only 0.3 K, making categorical temperature forecasts very error prone. For comparison, over the ENSO region, the terciles are 1.6 K apart. The second reason is that, in relative terms to the interannual variability, the error growth in the tropics is larger than in the extratropics, since the errors are mainly associated with convection rather than baroclinic instability as in the ex-
tratropics (Shukla 1981). This, together with low daily fluctuations, leads to a shorter upper limit of predictability.

In W2, prediction skill is substantially reduced almost everywhere. This particularly applies for continental skill, which now only more assumes values on the order of 0.1. Southern America and Africa thereby reveal slightly better prediction skill than North America, Eurasia, and Australia. Considerably more skill is found over the oceans, but also here the decrease w.r.t. W1 is pronounced.

Both W3 and W4 are characterized by the loss of basically all skill over the continents except for some areas in the tropics, in particular the northern part of South America, and Indonesia. Over the oceans, the ENSO region retains its high predictability well into W4. Generally, the patterns of skill distribution observed in W2 remain, but their magnitudes decrease.

It is important to note that there are only very few regions (e.g., the tropical Atlantic) where skill is negative, meaning that the forecasts are only rarely worse than climatology. This is in clear contrast to seasonal forecasts, where because of ensemble overconfidence, wide areas of negative RPSS<sub>2</sub> skill are found (Weigel et al. 2008). A further discussion of this aspect is provided in section 5.

So far, the focus has been on annual averages of MOFC prediction skill. We continue with an evaluation of the seasonal variability in prediction skill.

2) Seasonal and regional dependence of skill

Figure 5 displays the annual cycle of MOFC temperature prediction skill, averaged over all grid points of the northern and southern extratropics as well as the tropics. The time series are smoothed by applying a centered five-point moving-average filter with symmetric weights (1/9, 2/9, 3/9, 2/9, 1/9) (e.g., Chatfield 2004) to identify the major intra-annual characteristics. In all panels, the absolute skill drop between forecast weeks W1 and W2 is seen to be independent of the time of the year, but not independent of the region. Indeed, the loss of skill is less pronounced in the tropics where, as has been discussed with Fig. 4, skill is either already low from the beginning (tropical Indian and Atlantic Ocean), or remains high for all lead times (ENSO region in the equatorial Pacific).

Over the northern extratropics, the minimum in W1 skill is found in boreal summer, which is consistent with common experience in conventional weather forecasting. The skill of W2, W3, and W4, on the other hand, does not reveal a significant annual cycle. If anything, W2 skill is seen to be slightly enhanced in late winter (February and March), a result that is consistent with the findings of Vitart (2004). A similar, but more pronounced, pattern of seasonal skill variability is found in the southern extratropics, where even W4 predictability reveals a significant annual cycle with a minimum in late boreal summer and a maximum in boreal winter. Over the tropics, on the other hand, this cycle is less pronounced, but still visible. The reasons for this enhanced winter predictability are not evaluated here, but they may well be linked with the magnitude of ENSO-induced SST anomalies, which are strongest in boreal winter.

To obtain more regional information of the prediction skill for different times of the year, global maps of seasonally averaged W2-skill of temperature have been calculated and are displayed in Figs. 6a–d for boreal winter, spring, summer, and autumn, respectively. The corresponding maps of W3 and W4 prediction skill are not shown here but can be found in Baggenstos (2007). They look qualitatively similar to Fig. 6, but with strongly reduced magnitude. Figures 7 and 8 display W2 prediction skill in greater detail for two economically important extratropical regions, namely, North America and Europe. Finally, in Table 1, the average prediction skill as well as its uncertainty have been calculated in dependence of season and lead time, averaged over the entire globe, the Niño-3.4 region (5°S–5°N, 170°–120°W), and over the continental climate regions as defined by Giorgi and Francisco (2000).

The figures and table are not discussed individually, but a summary and synthesis of the key results on MOFC temperature prediction skill is presented in the following.

1) The figures for global mean temperature prediction skill in the first row of Table 1 confirm the conclusions drawn above from Fig. 5, namely, that global predictability is stronger during boreal winter and weakest in boreal summer for all lead times considered.

2) The skill reveals a very pronounced seasonal dependence in the eastern and tropical Pacific, as can be seen from Fig. 6. The skill is seen to be stronger in autumn and winter, while being weaker in spring and summer. This is consistent with the skill values for the Niño-3.4 region in Table 1, which, for example, show a W2 skill of 0.67 ± 0.10 in winter, but only 0.34 ± 0.14 in spring.

3) Over most continental regions, the skill of W3 and W4 forecasts is at best marginal, but usually not even significant. Exceptions are some tropical regions such as the Amazon basin during all seasons, southern South America in spring, eastern and west-
FIG. 5. Annual cycle of average skill in (a) the northern extratropics (30°–85°N), (b) the southern extratropics (30°–85°S), and (c) the tropics (30°S–30°N). A five-point symmetric moving-average filter has been applied as described in the text. Both land and sea points are considered. A few typical confidence intervals are plotted to illustrate the range of uncertainty of the skill values obtained.
ern Africa in winter, and southern Africa in spring, revealing W3 and W4 skill on the order of 0.1.

4) The lowest value of continental W1 prediction skill is found over Southeast Asia during all seasons and over tropical Africa and South Asia in summer (on the order of 0.1–0.2). The reason is probably the enhanced error growth rate and the reduced interannual variability as discussed in section 4b(1).

5) The highest value of continental W1 and W2 skill is found in winter over eastern North America, with a skill value of 0.50 ± 0.05 and 0.25 ± 0.04, respectively. The enhanced W2 winter predictability over

![Fig. 6](image_url)  
**Fig. 6.** Seasonally averaged RPSS$_D$ skill of all W2 forecasts available for (a) boreal winter, (b) spring, (c) summer, and (d) autumn.

![Fig. 7](image_url)  
**Fig. 7.** As in Fig. 6, but for the North American domain.
the eastern United States is also clearly visible in Fig. 7a. Note that a different picture is obtained with the National Centers for Environmental Prediction (NCEP) medium-range model where maximum W2 winter skill is found in the upper Midwest (Hamill et al. 2004). During the rest of the year, North American W2 skill is much lower in spring and weakest in summer and autumn. No pronounced W3 prediction skill is found over any of the North American Giorgi regions (see Table 1), contrasting with the results of Vitart (2004) who found predictability in North America even beyond day 18.

6) Also in the European domain, W1 and W2 prediction skill tends to be higher in winter as compared with the other seasons. However, the seasonal variability in prediction skill is less pronounced than in North America and not always significant. Higher skill is observed if also ocean grid points are considered. For example, Fig. 8 shows that W2 prediction skill of the Mediterranian Sea is more skillful in autumn and winter, while over the North Sea exactly the opposite is found with a pronounced maximum in summer. As over most other extratropical continental regions, the W3 and W4 prediction skill of mean near-surface temperature is very low.

c. Persistence forecasts

The evaluations presented above are based on climatology as a reference with which to be compared. Another common reference strategy, which is sometimes even considered to be a harder test, is persistence. To evaluate the prediction skill of persistence forecasts, we follow the approach of Vitart (2004) and persist W1 probability forecasts to “predict” weeks W2, W3, and W4. In other words, the W2 (W3, W4) persistence forecast is a “normal” W1 probability forecast that is verified against the observations of W2 (W3, W4). The verification is carried out in the same way as described above (i.e., by applying the debiased RPSS$_D$ skill metric with climatology as a reference).

Maps of the annually averaged persistence prediction skill are shown in Fig. 9 for forecast weeks W2–W4. Already in W2, persistence performs worse than climatology on all continental landmasses except for equatorial South America. Over the oceans, significant positive skill is only observed over the ENSO region and parts of the Atlantic. A comparison with Fig. 4 reveals that it is only in these oceanic regions that persistence performs almost equally well as the MOFC predictions. The observed contrasts become even more pronounced during forecast weeks W3 and W4: only the ENSO region over the Pacific and the trade wind zones over the Atlantic retain their positive skill, indicating that the high MOFC skill observed in these regions (see Figs. 4c,d) is to a large degree a consequence of SST persistence effects. On the other hand, basically all continental areas as well as the extratropical oceans are characterized by significantly negative persistence skill. This is mainly because ensemble spread usually widens as the lead time increases (i.e., the W1 forecasts are typically less dispersed than forecasts for W2 and later; Vitart 2004). This implies that W1 ensemble distributions tend to be overconfident given the weak predictability found in forecast weeks W2–W4. Overconfidence, in turn, is a property that is heavily penalized by the RPSS$_D$ skill metric as has been shown by Weigel et al. (2008).

All in all, the results show that, in a probabilistic prediction context, and for these time scales, persistence is not a competitive alternative for MOFC temperature forecasts or even climatology (over the continents).

5. Discussion

It is not the aim of this study to evaluate the physical reasons behind the observed spatial and seasonal variability in prediction skill. Rather, in our discussion we...
TABLE 1. Local skill in dependence of season and forecast week averaged over the entire world (first row, considering both land and sea grid points), the Niño-3.4 region (second row), as well as the 21 climate regions (only land points) defined by Giorgi and Francisco (2000). The numbers in brackets represent the half-width of the 5%–95% confidence intervals of the skill estimates. Months are abbreviated by first letter of each month [e.g., December–February (DJF)].

<table>
<thead>
<tr>
<th></th>
<th>DJF</th>
<th>MAM</th>
<th>JJA</th>
<th>SON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W1</td>
<td>W2</td>
<td>W3</td>
<td>W4</td>
</tr>
<tr>
<td>World</td>
<td>38 (2)</td>
<td>18 (2)</td>
<td>12 (2)</td>
<td>9 (2)</td>
</tr>
<tr>
<td>Niño-3.4</td>
<td>70 (9)</td>
<td>67 (10)</td>
<td>64 (10)</td>
<td>64 (10)</td>
</tr>
<tr>
<td>Australia</td>
<td>27 (3)</td>
<td>5 (4)</td>
<td>4 (2)</td>
<td>5 (2)</td>
</tr>
<tr>
<td>Amazon basin</td>
<td>22 (5)</td>
<td>16 (6)</td>
<td>11 (7)</td>
<td>10 (7)</td>
</tr>
<tr>
<td>Southern South America</td>
<td>22 (3)</td>
<td>10 (4)</td>
<td>4 (5)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Central America</td>
<td>33 (4)</td>
<td>17 (4)</td>
<td>4 (4)</td>
<td>7 (3)</td>
</tr>
<tr>
<td>Western North America</td>
<td>39 (5)</td>
<td>12 (4)</td>
<td>4 (4)</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Central North America</td>
<td>48 (4)</td>
<td>18 (4)</td>
<td>4 (5)</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Eastern North America</td>
<td>50 (5)</td>
<td>25 (4)</td>
<td>5 (6)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Alaska</td>
<td>40 (6)</td>
<td>14 (6)</td>
<td>-1 (5)</td>
<td>-5 (3)</td>
</tr>
<tr>
<td>Greenland</td>
<td>38 (3)</td>
<td>12 (3)</td>
<td>5 (3)</td>
<td>3 (4)</td>
</tr>
<tr>
<td>Mediterranean</td>
<td>44 (2)</td>
<td>12 (4)</td>
<td>7 (4)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Northern Europe</td>
<td>44 (3)</td>
<td>11 (7)</td>
<td>5 (5)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Western Africa</td>
<td>30 (5)</td>
<td>16 (3)</td>
<td>10 (4)</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Eastern Africa</td>
<td>28 (5)</td>
<td>13 (4)</td>
<td>11 (4)</td>
<td>8 (3)</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>25 (4)</td>
<td>10 (5)</td>
<td>4 (4)</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Sahara</td>
<td>40 (4)</td>
<td>14 (2)</td>
<td>8 (3)</td>
<td>7 (3)</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>18 (6)</td>
<td>9 (6)</td>
<td>8 (6)</td>
<td>7 (7)</td>
</tr>
<tr>
<td>East Asia</td>
<td>39 (6)</td>
<td>11 (7)</td>
<td>4 (5)</td>
<td>3 (5)</td>
</tr>
<tr>
<td>South Asia</td>
<td>33 (4)</td>
<td>13 (4)</td>
<td>9 (2)</td>
<td>4 (3)</td>
</tr>
<tr>
<td>Central Asia</td>
<td>47 (4)</td>
<td>17 (5)</td>
<td>5 (5)</td>
<td>6 (4)</td>
</tr>
<tr>
<td>Tibet</td>
<td>39 (3)</td>
<td>13 (5)</td>
<td>5 (4)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>North Asia</td>
<td>43 (3)</td>
<td>13 (4)</td>
<td>4 (3)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>
want to focus on a key result, which is most relevant from a user perspective; namely, the finding that extratropical continental prediction skill essentially vanishes after forecast week W2 (i.e., that the MOFC does not seem to provide additional value with respect to climatology for lead times that are beyond the prediction range of medium-range weather forecasts). In this context, we want to discuss the following two questions: Could the limit of predictability be extended beyond W2 by applying appropriate postprocessing methods? Could the limit of predictability be extended by increasing the averaging period? Apart from that, we want to discuss an example that shows that, even if the average prediction skill is low, there may nevertheless be some forecasts that contain useful information.

a. Can recalibration improve the monthly forecasts beyond forecast week W2?

For most probabilistic skill scores, including the RPSS$_D$, zero skill as observed in forecast weeks W3 and W4 can be because of two reasons: (i) the absence of a physically predictive signal per se, or (ii) wrong ensemble spread (i.e., poor reliability). In the latter case, some improvement in prediction skill could at least in principle be achieved a posteriori by recalibration [Toth et al. (2003), e.g., by inflating the ensemble spread as for example described in Doblas-Reyes et al. (2005)]. To estimate the potentially achievable skill, we apply a method published in Hamill et al. (2004). By random sampling from bivariate normal distributions having specified correlation, Hamill et al. (2004) generate correlated time series of synthetic observations and ensemble means. From these, they establish an empirical relationship between the expected RPSS of reliable forecasts on the one hand and the correlation on the other hand (see their Fig. 5d). To investigate whether the MOFC forecasts follow this curve, the correlation coefficients between the MOFC ensemble mean forecasts and the corresponding observations have been calculated for each week of the year and each grid
point for forecast weeks W1–W4. Maps of the resulting annually averaged correlation coefficients are shown in Fig. 10. In Fig. 11, all annually averaged continental \( \text{RPSS}_D \) values of Fig. 4 are plotted against the corresponding correlation values of Fig. 10 for the four lead times. The method of Hamill et al. (2004) has been adjusted for finite-sized ensembles, and the resulting \( \text{RPSS}_D \) correlation curve is shown in Fig. 11 as a gray line. Assuming Gaussian behavior of the observed and simulated weekly temperature averages, this relationship essentially displays the \( \text{RPSS}_D \) values that would be obtained were the MOFC forecasts reliable. Admittedly, the Gaussian assumption is a very simplifying one, but it can be justified as a first rough estimate for the variable considered (Wilks 2002, 2006). Figure 11 shows that particularly during W1 many points are below the gray line, implying that the prediction skill can be potentially improved by recalibration. However, during W3 and particularly W4 the MOFC prediction skill follows the gray line (i.e., the continental W3 and W4 forecasts are essentially reliable and can hardly be further improved by recalibration).

While this result may appear disappointing, it also has a positive facet in that the \( \text{RPSS}_D \) is generally not negative. In other words, in the worst case MOFC forecasts are, due to their reliability, essentially identical to the climatological forecast, but not worse. This implies that the use of MOFC forecasts, while not always being valuable, is generally not harmful either. This contrasts with longer-term seasonal forecasts, which often do reveal negative skill due to strong overconfidence and can therefore be worse than climatological guessing (Weigel et al. 2008).

**b. Can larger averaging periods improve prediction skill?**

In all our analyses so far, we have only considered 7-day temperature averages. The question rises whether predictability can be increased and the limit of pre-
dictability extended by choosing another averaging period. We evaluate this question for two example regions, namely, Europe and the Niño-3.4 region. For these regions, annually averaged prediction skill is now not only calculated for 7-day temperature averages, but also for 3- and 14-day averages. The results are displayed in Fig. 12 as a function of the number of days between initialization and the end of the respective prediction interval. Note that the lead time is now incremented by day rather than by week as before. The picture shows that both for Europe and the Niño-3.4 region skill increases as the length of the averaging interval becomes larger. This is consistent with the study of Rodwell and Doblas-Reyes (2006) who evaluated the effect of averaging period on temperature forecasts during the European heat wave of 2003. There are two reasons for this behavior. First, increasing the averaging period implies that the higher skill of earlier lead times contributes to the averages. For example, the 3-day prediction interval that ends 20 days after initialization covers the forecasting days 18–20, while the corresponding 7-day (14 day) interval includes days 14–20 (7–20; i.e., contains more of the high predictability found for shorter lead times). Particularly in Europe this can have a big effect, where skill is observed to drop rapidly after only one week from initialization. Second, nonpredictable high-frequency noise is filtered out as the prediction intervals grow. Particularly for the Niño-3.4 region, the latter effect seems to be dominant, since skill drops too slowly with lead time as to explain the observed gain in prediction skill solely by the inclusion of days with shorter lead and higher skill. The observation that skill can be enhanced by increasing the averaging intervals is certainly a result that is not very surprising; on the contrary, it may appear rather trivial. However, this property is only rarely explicitly mentioned, even though it may have quite some relevance from a user perspective: it implies that the limit of pre-

![Fig. 11. Annually averaged RPSS\(_D\) plotted against annually averaged ensemble mean correlation for all continental grid points for forecast weeks (a) W1, (b) W2, (c) W3, and (d) W4. The gray lines show the RPSS\(_D\) values that would be obtained were the forecasts reliable.](image)
dictability can be significantly extended into the future for users who do not require forecast information in temporal resolutions as high as weekly. The same applies for users who feed raw forecasts into application models with some degree of memory or inertia (i.e., with integrating characteristics that carry the high prediction skill of short-lead forecasts further into the future). For instance, preliminary results from a case study where MOFC forecasts over Switzerland have been fed into a soil moisture prediction model, have revealed that the limit of predictability is extended by almost a week if the output of the soil moisture predictions is evaluated rather than the raw MOFC forecasts (Bolius et al. 2007).

c. A further comment

From seasonal forecasting, it is known that prediction skill can be closely linked to the presence or absence of certain climatic boundary conditions. For example, seasonal winter prediction skill in the North American–Pacific region is higher during ENSO years than during the other years (Shukla et al. 2000). Similarly, it may be possible to find situations of enhanced predictability on the monthly time scale beyond W2, even if the average predictability is low. As an example, we consider MOFC winter forecasts (December–March) for temperature in continental Europe. We stratify the verification samples on two subsets, depending on whether the prediction week falls into a phase of positive or negative North Atlantic Oscillation (NAO) index [the NAO index is defined as the sea level pressure difference between the Azores and Iceland and is closely connected to the storminess characteristics in Europe (Hurrell 1995; Jung et al. 2003); data of the Climate Research Unit (CRU) Web site page have been used (see online at http://www.cru.uea.ac.uk/cru/data/nao.htm)]. These two subsets are verified separately and displayed in Fig. 13 as a function of prediction time (i.e., the time between initialization and the end of the prediction interval), again with lead time being incremented daywise as in Fig. 12a rather than weekwise as before. The results show that for prediction times of 12–28-day skill during negative NAO (NAO⁻) phases is higher than during positive NAO (NAO⁺) phases. The reasons for this observation may be linked to the augmented occurrence of persistent blockings observed during NAO⁻ phases (e.g., Pavan et al. 2000; Scherrer et al. 2006), but are not further investigated here. To quantify the significance of this finding, the 10% and 90% percentiles of the skill difference between NAO⁻ and NAO⁺ phases have been estimated by bootstrapping (cf. section 3c). The skill difference and the corresponding confidence range is plotted in Fig. 13b, showing that predictability during NAO⁻ phases outperforms NAO⁺ skill on the 90% confidence level at least for prediction times from 15 to

![Image](https://example.com/image.png)
19 days. Of course, in practical terms this result is not very useful, since there is no value in predicting the predictability when the forecasts are reliable anyway. This example should rather be understood as a demonstration that, even when the average predictability is close to zero, there may be climatological conditions under which the skill is enhanced and the forecasts actually do contain more useful information than suggested by the averaged figures in Table 1. This conclusion is probably not only restricted to the presence or absence of certain flow regimes, but may also apply to other aspects such as the surface conditions. For example, it is imaginable that the MOFC predictability of the European summer may depend on the amplitude of the initial soil moisture anomalies.

All in all, the aforementioned decline of MOFC prediction skill for lead times beyond forecasting week W2 would be less stringent than suggested by Table 1, if larger averaging intervals were considered. However, the value of increased averaging periods is highly user specific and to some degree only hypothetical. Because of the surprisingly good reliability, which by itself is a very favorable property, it is not expected that any form of recalibration would significantly improve the MOFC prediction skill.

6. Conclusions

In this paper we presented a probabilistic verification of surface temperature forecasts obtained from...
the operational ECMWF monthly forecasting system, a coupled atmosphere–ocean global circulation model. The main objectives of this study have been to (i) define and set up an appropriate probabilistic verification framework, and (ii) to apply this verification framework for a systematic and fully probabilistic skill assessment of MOFC temperature forecasts in dependence of lead time, region, and season. The verification is based on the debiased ranked probability skill score (RPSS), a probabilistic and strictly proper skill score that has the advantage of not penalizing the intrinsic unreliability induced by small ensemble sizes—a property that is favorable in the present context, since the ensemble size of MOFC hindcasts is very small. The RPSS has been generalized such that the small-sized hindcast ensembles and large-sized forecast ensembles can be jointly considered in the verification. Moreover, in contrast to earlier work of this type, the sampling uncertainty of the skill values obtained has been estimated by a bootstrapping approach. The following results have been found.

1) Already from the first forecasting week, the MOFC reveals significant systematic biases on the order of 1 K and more, also over land. While this result is not surprising, given that no measures are applied to suppress model drift, it manifests the need for hindcasts to estimate the model climatology and to correct for such systematic errors.

2) The seasonal cycle of global predictability reveals higher skill in boreal autumn and winter and lower skill in boreal spring and summer for all lead times.

3) Continental prediction skill during forecast week W1 (days 5–11 from initialization) is on the order of 0.4–0.6 in the extratropics and on the order of 0.2–0.3 in the tropics. In forecast week W2 (days 12–18), continental skill drops to much lower values of about 0.1–0.2. The highest continental W1 and W2 prediction skill is found over the eastern United States in boreal winter.

4) Beyond forecast week W2, pronounced prediction skill is predominantly found over the oceans, particularly over the ENSO region in the central and eastern Pacific. Continental skill, however, essentially vanishes after W2, apart from some areas in tropical Africa and South America.

5) An alternative forecasting strategy that is based on persisting W1 forecasts turns out to be worse than climatological guessing, at least over land grid points and in a probabilistic verification context.

6) Only little can be done to improve continental prediction skill of near-surface temperature beyond forecast week W2 for the investigated model. The potential of a posteriori recalibration is limited because the forecasts are surprisingly reliable. Some skill improvement can be achieved by increasing the averaging period considered. While this averaging effect is to some degree trivial in that simply the high-skill forecasts of shorter lead times are carried further into the future, this may still be of relevance for specific user groups.

7) When the average prediction skill is low, there may nevertheless be climatologic conditions under which the forecasts contain some useful information. For example, there is evidence that during NAO+ conditions in the winter months, the limit of temperature predictability for continental Europe is extended by about 4–5 days.

All in all, the results presented may appear disappointing in that they imply that the MOFC does not have much additional information on time scales that are not already covered by (slightly extended) medium-range weather forecasts. While only temperature forecasts have been considered here, we do not expect fundamentally better results for other variables such as precipitation, which is often considered to be even harder to predict. However, the observed low prediction skill of tercile-based weekly averages does not necessarily imply an equally low prediction skill of extreme events, and it could be worthwhile to further investigate this aspect.

Generally, we believe that a major improvement in monthly predictability might be feasible once the Madden–Julian oscillation (MJO), the most important source of intraseasonal variability, is better represented by the model, at least in the tropics. Having said that, there are indications that the MJO might even affect the intraseasonal variability in the extratropics (Ferranti et al. 1990; Cassou 2008). Currently, however, the MOFC is not able to predict the MJO beyond 14 days and has problems maintaining the amplitude of an MJO event. Vitart et al. (2007) stress that better parameterizations and a better representation of ocean mixing is needed. This, together with a better representation of other slowly evolving boundaries the troposphere interacts with (e.g., the sea surface, soil moisture, snow cover and the stratosphere) might eventually affect some improvement in MOFC prediction skill.

Acknowledgments. This study was supported by the Swiss National Science Foundation through the National Center for Competence in Research (NCCR–Climate) and by the ENSEMBLES project (EU FP6 Contract GOCE-CT-2003-505539).
APPENDIX

Effective Ensemble Size

Weigel et al. (2007a) have shown that the RPSSD for ensemble forecasts with ensemble size M is derived by replacing the $\langle \text{RPS}_{\text{Cl}} \rangle$ in Eq. (1) with the expectation $\epsilon$ of the scores $\text{RPS}_{\text{ran},M'}$, which an $M$-member ensemble prediction system would produce in the case of merely random climatological resamples:

$$\text{RPSS}_D = 1 - \frac{\langle \text{RPS} \rangle}{N_h + N_f} \text{RPS}_{\text{ran},M'} + \frac{N_f \text{RPS}_{\text{ran},M'}}{N_h + N_f}.$$

(A1)

Assume that a skill score is now calculated on the basis of $N_h$ hindcasts with ensemble size $M_h$, and $N_f$ forecasts with ensemble size $M_f$. This composition must then be equally represented in the reference:

$$\text{RPSS}_D = 1 - \frac{\langle \text{RPS} \rangle}{N_h + N_f} \text{RPS}_{\text{ran},M_h} + \frac{N_f \text{RPS}_{\text{ran},M_f}}{N_h + N_f}.$$

Applying Eqs. (10)–(15) of Weigel et al. (2007a), together with Eq. (3) of the present paper, one obtains the following:

$$\text{RPSS}_D = 1 - \frac{\langle \text{RPS} \rangle}{N_h + N_f} \text{RPS}_{\text{Cl}} + \frac{N_f \text{RPS}_{\text{Cl}}}{N_h + N_f} + \frac{N_f}{N_h + N_f} \langle \text{RPS} \rangle.$$

$$= 1 - \frac{\langle \text{RPS} \rangle}{\langle \text{RPS}_{\text{Cl}} \rangle + D_0/M_{\text{eff}}}.$$

with

$$M_{\text{eff}} = M_h M_f / (M_h N_f + M_f N_h).$$

REFERENCES


ECMWF, cited 2007: Monthly forecasting. [Available online at


