Monthly Forecasting at ECMWF

FRÉDÉRIC VITART
European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, United Kingdom

(Manuscript received 9 March 2004, in final form 25 May 2004)

ABSTRACT

A monthly forecasting system based on 32-day coupled ocean–atmosphere integrations has been set up at ECMWF. This system has run routinely since March 2002 every 2 weeks, and 45 cases from March 2002 to December 2003 have been verified. Results of this validation suggest that the model displays some skill in predicting weekly averaged 2-m temperature, precipitation, and mean sea level pressure anomalies relative to the climate of the past 12 years. For days 12–18, probabilistic scores indicate that the monthly forecasting system performs generally better than both climatology and the persistence of the previous weekly probabilities, suggesting that forecasts at that time range could be useful. After about 20 days of forecast, the model displays some skill in predicting events with a large threshold. At that time range, the performance of the system depends strongly on the geographical location, with Europe being a particularly difficult region. However, the model displays some useful skill after 20 days over North America, Asia, and the southern extratropics.

In order to calibrate the monthly forecasting system, a 5-member hindcast over the 12 years preceding the real-time forecast has been produced. Probabilistic scores computed with the hindcast confirm the main results obtained with the real-time forecast. The scores display strong seasonal variability, with the model being particularly skillful in winter.

1. Introduction

Medium-range weather forecasting is essentially an atmospheric initial value problem. Sea surface temperatures (SSTs) are usually simply persisted. Seasonal forecasting, on the other hand, is justified by the long time scale of the predictability of the oceanic circulation (of the order of several months) and by the fact that the tropical SSTs have a significant global impact on the atmospheric circulation. Since the oceanic circulation is a major source of atmospheric predictability at seasonal time scales, the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting system is based on coupled ocean–atmosphere integrations.

Monthly forecasting (forecasts between 10 and 30 days) fills the gap between medium-range weather forecasting and seasonal forecasting. It is often considered a difficult time range for weather forecasting, since the time scale is sufficiently long so that much of the memory of the atmospheric initial conditions is lost, and it is probably too short so that the variability of the ocean is not large enough, which makes it difficult to beat persistence. However, an important source of predictability at this time range is the Madden–Julian oscillation (MJO); (see, e.g., Ferranti et al. 1990). Another potential source of predictability includes the stratospheric initial conditions that can project in the troposphere over a time scale of a month (Baldwin et al. 2003).

The interest in monthly forecasting was triggered by Miyakoda et al. (1983). This paper showed how the pronounced blocking event of 1977 was successfully reproduced in 1-month forecasts produced by some general circulation models. Miyakoda et al. (1986) found some marginal skill in eight January 1-month integrations using a 10-day mean filter applied to the prognoses. The report of successful forecasts beyond day 10 triggered a lot of interest at that time. Many of the world’s operational prediction centers started to produce a number of large experiments on extended-range forecasting (Tracton et al. 1989; Owen and Palmer 1987; Molteni et al. 1986; Déqué and Royer 1992). ECMWF used its operational forecast to produce a pair of 31-day forecasts starting at two consecutive days for every month from April 1985 to January 1989 (Palmer et al. 1990). These experiments generally showed some moderate skill after 10 days (Miyakoda et al. 1986; Déqué and Royer 1992; Brankovic et al. 1988), particularly when comparing the forecast to climatology. However, a particularly tough test for extended-range forecasting is to beat persistence. At ECMWF, the extended-range experiments described in Molteni et al. (1986) failed to produce forecasts after 10 days that were significantly better than persisting the medium-range operational
forecasts. As a consequence, this experiment did not lead to an operational extended-range forecasting system at ECMWF. Anderson and Van den Dool (1994) added an additional pessimistic note to this problem, demonstrating that some apparent high quality forecasts in the extended range that triggered the initial enthusiasm for monthly forecasting could just occur by chance. Using the extended-range model from National Centers for Environmental Prediction (NCEP) Dynamical Extended-Range Forecasting (DERF) (Tracton et al. 1989), they found that after 12 days, the model did not produce better forecasts than a no-skill control. However, Newman et al. (2003) found some strong predictability of week 2 and week 3 averages in some regions of the Northern Hemisphere using a statistical linear inverse model (LIM).

The main goal of the present paper is to describe the ECMWF monthly forecasting system and revisit the issue of monthly forecasting by evaluating the skill of a state-of-the-art general circulation model (GCM) in the extended time range. Since the 1980s, medium-range forecasting has made considerable progress, with about 2 days of predictability gained in the last 20 yr, because of improvements in the model physics and improvements in the generation of initial conditions. This is likely to have beneficial impact for the extended range. The GCM used in this study has a much finer resolution (T159L40) than the models used in the 1980s, although Tibaldi et al. (1988) did not notice any significant improvement in the extended-range scores when increasing the horizontal resolution. Finally, the ECMWF monthly forecasting system is based on fully coupled ocean–atmosphere integrations rather than on atmospheric-only integrations forced by persisted SSTs. The ocean–atmosphere coupling may help to capture some aspects of the MJO variability (Flatau et al. 1997). The ECMWF monthly forecasting system has also a much larger ensemble size (51 members) than in any previous studies, which helps to produce more accurate probabilistic forecasts. The verification of monthly forecasting in the 1980s was mostly focusing on anomaly correlation and root-mean-square (RMS) error scores of the ensemble mean. However, after 10 days, the forecasts are no longer deterministic, and the ensemble mean is not necessarily the best representation of the forecast, particularly in the case of multimodality of the ensemble distribution. Therefore, the present paper focuses mostly on probabilistic scores.

The ECMWF monthly forecasting system is described in section 2 of the present paper. Deterministic and probabilistic scores of the 51-member ensemble real-time forecast are discussed in section 3. Scores obtained with a five-member hindcast over 12 yr are displayed in section 4. Section 5 concludes this study, along with a discussion.

2. Description of the monthly forecasting system

The monthly forecasts are based on an ensemble of 51 coupled ocean–atmosphere integrations (one control and 50 perturbed forecasts). The length of the coupled integration is 32 days, and the frequency of the monthly forecasts is currently every 2 weeks. The atmospheric component is the ECMWF atmospheric model Integrated Forecast System (IFS), using the same model version as is used to produce the ECMWF operational forecast. Currently, the atmospheric model is run at T159 resolution (1.125° × 1.125°) with 40 levels in the vertical. This represents a resolution in between the Ensemble Prediction System (EPS) (T255L40) and seasonal forecasting (T95L40). The oceanic component is the same as for the current ECMWF seasonal forecasting system. It is the Hamburg Ocean Primitive Equation (HOPE) model (Wolff et al. 1997). The ocean model has lower resolution in the extratropics but higher resolution in the equatorial region, in order to resolve ocean baroclinic waves and processes that are tightly trapped at the equator. The ocean model has 29 levels in the vertical. The atmosphere and ocean communicate with each other through the Ocean Atmosphere Sea Ice Soil (OASIS) coupler (Terray et al. 1995). The atmospheric fluxes of momentum, heat, and freshwater are passed to the ocean every hour.

Atmospheric and land surface initial conditions are obtained from the ECMWF operational atmospheric analysis/ reassessment system. Oceanic initial conditions originate from the oceanic data assimilation system used to produce the initial conditions of the ECMWF seasonal forecasting system. However, this oceanic data assimilation system lags about 12 days behind real time. In order to “predict” the ocean initial conditions, the ocean model is integrated from the last analysis, forced by wind stress, heat fluxes and precipitation–evaporation (P–E) from the operational atmospheric analysis system. During this “ocean forecast,” the sea surface temperature is relaxed toward persisted SST, with a damping rate of 100 W m−2 K−1. This method allows us to produce monthly forecasts in “real time” without having to wait for the ocean analysis to be ready.

The monthly forecasting system is run 51 times from slightly different initial conditions. One forecast, called the control, is run from the operational ocean and atmosphere ECMWF analyses. The 50 additional integrations, the perturbed members, are made from slightly different initial atmospheric and oceanic conditions, which are designed to represent the uncertainties inherent in the operational analyses. The atmospheric component of the coupled GCM is perturbed in the same way as in the EPS for medium-range forecasts. The 50 perturbations are produced using the singular vector method (Buizza and Palmer 1995). These include perturbations in the extratropics and perturbations in some tropical areas by targeting tropical cyclones (Puri et al. 2001). In addition, in order to take account for the effects of uncertainties in the model subgrid-scale parameterizations, the tendencies in the atmospheric physics are randomly perturbed during the model integrations. The current implementation is the same as that used in...
EPS. For each ensemble member, the stochastic physics (Buizza et al. 1999; Palmer 2001) perturbs gridpoint tendencies of the physics up to 50%. The tendencies are multiplied by a random factor drawn from an uniform distribution between 0.5 and 1.5. The random factor is constant within a $10^5 \times 10^5$ domain for 6 h. The whole globe is perturbed.

The oceanic initial conditions are perturbed in the same way as in the operational ECMWF seasonal forecasting system (Anderson et al. 2003). A set of SST perturbations has been constructed by taking the difference between two weekly mean SST analyses from NCEP [Reynolds OIv2 and Reynolds two-dimensional variational data assimilation (2DVAR), both described in Reynolds et al. (2002)] from 1985 to 1999. A second set of SST perturbations has been constructed by taking the difference between Reynolds 2DVAR SSTs and its 1-week persistence. The first set of SST perturbations samples the uncertainties in the SST analysis, whereas the second difference samples the uncertainties due to the fact that the SSTs from NCEP are a weekly mean product. For each starting date, 25 combinations from these two different sets of perturbations are randomly selected and are added to the SSTs produced by the operational ocean analyses with a $+$ or $-$ sign, creating 50 perturbed initial states. In order to have a 3D structure, the SST perturbations are linearly interpolated from the full value at the surface to zero at an oceanic depth of 40 m. A set of wind stress perturbations is also calculated by taking the difference between two monthly wind stress analyses [the ECMWF 15-yr Re-Analysis (ERA-15)/ECMWF analysis and a wind analysis from Southampton Oceanography Centre described in Josey et al. (1998)] from 1980 to 1997. Five ocean assimilations (one control and four perturbed) are produced by randomly picking two perturbations from the set of wind stress perturbations for each month of data assimilation and adding them with a $+$ or $-$ sign to the analyzed wind stress. The same perturbations cannot be chosen for two consecutive months. The wind stress and SST perturbations are combined to produce the 50 perturbed oceanic initial conditions. Current research is exploring the possibility of using oceanic singular vectors to perturb the oceanic initial conditions instead of the method described above.

After 10 days of coupled integrations, the model mean climate begins to be different from analysis (see, e.g., Fig. 1). No "artificial" terms are introduced to try to reduce the drift of the model, and no steps are taken to remove or reduce any imbalances in the coupled model initial state; we simply couple the models together and integrate forward. The effect of the drift on the model calculations is estimated from previous integrations of the model in previous years (the back statistics). The drift is removed from the model solution during the postprocessing. In the present system, the climatology
Fig. 2. Probability of 2-m temperature anomaly predicted by the monthly forecasting system to be greater than 0. Each panel represents one 7-day period. The forecast starting date is 12 Mar 2003. The ensemble distribution of the real-time forecast has been compared to the distribution of the model climatology (hindcasts), and a WMW test has been applied in order to determine if the difference is significant. Areas where the significance is less than 90% have been blanked. The dotted lines indicate 95% significance with the WMW test.

(back statistics) is a 5-member ensemble of 32-day coupled integrations, starting on the same day and month as the real-time forecast for each of the past 12 yr. For instance, the first starting date of the real-time forecast was 27 March 2002. The corresponding climatology is a 5-member ensemble starting on 27 March 1990, 27 March 1991, ... , 27 March 2001. The 5-member ensemble is thus integrated with 12 different starting dates. This represents a total of 60 integrations and constitutes the 60-member ensemble of the back statistics. The back statistics are created every 2 weeks, alternately with the real-time forecast. Figure 1 displays the weekly evolution of the model drift of geopotential height at 500 hPa. During the first weekly period (days 5–11), the drift is relatively small, but increases regularly week by week. The pattern of the model bias in that time scale is consistent with the patterns of the model bias in the seasonal time range. Experiments where the atmosphere is forced by observed SSTs indicate that the model drift in the 30-day time range originates essentially from errors in the atmospheric circulation model. The contribution of the ocean–atmosphere coupling to the model drift is comparatively small during the first month of the coupled model integrations.

Monthly forecasting products at ECMWF include anomaly, probability, and tercile maps based on comparing the 51-member ensemble distribution of the real-time forecast to the 60-member ensemble distribution of the model climatology (hindcast). The forecasts are based on weekly means since it is likely that at that time range, the model has more skill in predicting weekly anomalies than daily values. Fields like surface temperature, 2-m temperature, precipitation, and mean sea level pressure have been averaged over 7 days. The 7-day periods correspond to days 5–11, days 12–18, days 19–25, and days 26–32. They have been chosen that way so that they correspond to Sunday to Saturday calendar weeks (the monthly forecasting starting date is on Wednesday at 0000 UTC). For the purpose of evaluating the skill of extended-range forecasts, this definition has the advantage that the second weekly period is beyond day 10 and corresponds almost to the first week after the 10-day time range. In a more classic definition of weeks, skill in the second week (days 8–14) would be more difficult to interpret, since part of the skill could be from the contribution of days 8 to 10. The length of the monthly forecasting system is 32 days, so that it contains four of these weekly periods. Figure 2 displays a typical example of a probability map produced by the ECMWF monthly forecasting system. The example dis-
played in Fig. 2 is the probability that the weekly mean 2-m temperature anomalies (relative to the model climatology from the past 12 yr) predicted by the monthly forecast starting on 12 March 2003 are positive. A Wilcoxon–Mann–Whitney test (WMW test; see, e.g., Wonacott and Wonacott 1977) has been applied to estimate whether the ensemble distribution of the real-time forecast is significantly different from the ensemble distribution of the back statistics. Regions where the WMW test displays a significance less than 90% are blank. In other words, shaded areas in Fig. 2 represent areas where the model displays some potential predictability. Unsurprisingly, the percentage of areas that are shaded decreases week by week, indicating that the model is drifting toward its climatology. In general, the model displays strong potential predictability over a large portion of the extratropics for days 12–18. However, there is generally a sharp decrease of potential predictability during the last two weeks of the forecasts. After 20 days of forecasts, the ensemble distribution is generally close to the model climatology. However, there have been several cases, like the forecast starting on 31 December 2003, where the monthly forecast displayed strong potential predictability in the last two weeks of the forecast over most of the Northern Hemisphere.

The ECMWF monthly forecasting system has been run routinely since 27 March 2002. The following sections discuss the verification of 45 real-time cases, between 27 March 2002 and 17 December 2003. The analysis used to verify the monthly forecasting system is the ECWMF operational analysis, or ERA-40 reanalysis (Uppala 2001) when available. For precipitation, the operational or ERA-40 forecasts of precipitation between 12 and 36 h were used as verification data.

3. Verification scores

a. ACC and RMS scores

In this section the performance of the monthly forecasting system is assessed by two skill scores: the correlation coefficients of the geopotential height anomalies (deviation from the climatological norms) between the ensemble mean of the forecasts and analysis, and the root-mean-square (RMS) error of the geopotential height at 500 hPa. As discussed in the introduction, these scores are probably not the most suitable for assessing probabilistic forecasts after 10 days, but they are the most widely used for evaluating medium-range weather forecasting and they were used in previous papers on monthly forecasting (see, e.g., Miyakoda et al. 1986; Déqué and Royer 1992; Brankovic et al. 1988).

For all 45 cases, the anomaly correlation and RMS scores of the ensemble mean were calculated. The scores obtained during the first 10 days of the forecast are slightly lower than those obtained with the operational ECMWF EPS, but the difference is small (not shown). Figure 3 displays the scores of the ensemble mean for all the 45 cases over the Northern Hemisphere extratropics (north of 30°N). Each line represents one individual case, and the dark line represents the mean over the 45 cases. Murphy and Epstein (1989) have shown that forecasts with an anomaly correlation larger than 0.5 are skillful relative to climatology. An anomaly correlation of 0.5 was reached between days 7 and 21, and on average at day 10. The linear correlation diminished quite sharply after day 10 and reached 0.3 around day 13 on average. The RMS error of the ensemble mean (full line in bottom panel of Fig. 3) reached climatology around day 15, suggesting that there was little deterministic skill in the monthly forecasting system after about 15 days of forecast.

There was some variability depending on the geographical area. For instance, over Europe, the anomaly correlations decreased much quicker than over the Northern Hemisphere extratropics as a whole. The RMS error over the North Pacific reached climatology around day 18, which was almost one week later than over Europe. There was also some seasonal variability, with summer being a particularly difficult season.

The results presented in Fig. 3 are based on daily
values. Miyakoda et al. (1986) used a 10-day running mean for both the forecasts and the observed data in an attempt to obtain positive skills by filtering out the unpredictable components, most especially the baroclinic eddies. In the present paper, we consider weekly mean periods instead, since the products of the ECMWF monthly forecasting system are based on weekly forecasts (section 2). The anomaly correlations and RMS errors were calculated on this weekly mean basis (Fig. 4). The anomaly correlation was of the order of 0.75 for days 5–11, about 0.35 for the days 12–18, about 0.2 for the days 19–25, and 0.17 for the days 19–32. The RMS error plot suggests that the RMS error of the monthly forecasting system reached climatology by days 12–18. The anomaly correlations for the three last weekly periods were significantly higher than those obtained by persisting the anomalies of the previous week (gray curve in Fig. 4). This suggests that although the deterministic skill of the monthly forecasting system is low after 10 days of forecasts, it is still higher than persistence. It is also higher than climatology for anomaly correlations. This suggests that the monthly forecasting system may be useful at that time range.

b. Probabilistic scores

In this section, the probabilistic scores of the monthly forecasting system will be evaluated through the scores obtained with weekly averaged surface temperature, 2-m temperature, precipitation, and mean sea level pressure.

The relative operating characteristics (ROC) scores are based on contingency tables giving the number of observed occurrences and nonoccurrences of an event as a function of the forecast occurrences and nonoccurrences of that event. The events are defined as binary, for instance, the probability that 2-m temperature is in the upper tercile. Hit rates and false alarm rates are computed for each probability interval (N bins), giving N points on a graph of hit rate (vertical axis) against false alarm rate (horizontal axis). The curve by definition passes through the points (0, 0) and (1, 1). The farther the curve lies toward the upper-left corner, the better. No skill forecasts are indicated by a diagonal line (hit rate = false alarm rate). The ROC score (Stanski et al. 1989; Mason and Graham 1999; Kharin and Zwiers 2003) is the area below the ROC curve. The closer the ROC score is to 1, the better. A no-skill forecast has an ROC score of 0.5.

Figure 5a displays an example of ROC diagrams obtained with four different periods: days 5–11, days 12–18, days 19–25, and days 26–32. In this example, the event scored is the probability that the surface temperature is in the upper tercile, over each grid point of the Northern Hemisphere extratropics. Only grid points over land were considered. For the monthly forecast, the upper tercile was computed relative to the model climatology. In that respect, the biases in the model ensemble mean and spread were taken into account. Figure 5 shows how the probabilistic scores decreased week by week. Over the first period (days 5–11), the ROC score was of order of 0.8, and dropped to 0.7 in the next week. It dropped again to about 0.6 in the following week. The ROC scores for days 19–25 and days 26–32 were close. The rate at which the ROC score degraded from one week to another depended on the variable considered. For precipitation (Fig. 5b), the ROC scores were lower than those obtained with the other variables for all the periods, and the ROC score for days 12–18 was about 0.6 only. For mean sea level pressure (Fig. 5c), the drop between days 5–11 to days 12–18 was larger than with other variables. There was also a significant drop between days 12–18 and days 19–25. For days 26–32, the ROC score was almost 0.5, which suggests that there was almost no skill at that time range for this variable. These results suggest that
for the period days 12–18, the model had some moderate skill, and performed better than climatology. For the two following weeks, the model displayed some low skill, but the performance seemed generally slightly better than climatology. In the rest of this section, the skill of the system for the periods days 12–18 and days 19–32 are investigated.

SKILL OF THE SYSTEM FOR THE PERIOD DAYS 12–18

The period days 12–18 corresponds almost to the first week after the end of the ECMWF medium-range forecast. At this time range the monthly forecasting system produces some significant signal in almost all forecast situations (see, e.g., Fig. 2, top-right panel). For precipitation, which is a much noisier field than 2-m temperature, areas with a significant signal were much smaller than with 2-m temperature.

In order to assess the skill of the monthly forecasting system at this time range, the ROC score was computed for each grid point over land on a 2.5° × 2.5° grid (Fig. 6) for the period March 2002–December 2003 (45 cases) and for the probability that 2-m temperature was in the upper tercile. According to Fig. 6, the ROC score exceeded 0.5 (shaded areas) in most regions. This suggests that, according to this test, the model produced better forecasts of probability that the 2-m temperature anomaly was in the upper tercile than climatology for the time range days 12–18. The ROC diagram computed over the 45 cases and over all the land points in the whole Northern Hemisphere extratropics (north of 30°N) confirmed the skill of the model relative to climatology, with an ROC score of 0.67, at least for this particular event and period of verification.

Another way of assessing the skill of probabilistic forecasts is the use of reliability diagrams (Wilks 1995). Figure 7a displays the attributes diagram of the probability that the 2-m temperature averaged over the days 12–18 is in the upper tercile. This graph displays the observed frequency as a function of the forecast probability. For a perfectly reliable forecast, the graph should lie along the 45° diagonal. For a no-skill forecast, the graph would be a horizontal line. According to Fig. 7a, the model forecasts seemed somewhat reliable. The observed frequency increased with higher forecast probabilities. The fact that the reliability graph in Fig. 7a was flatter than the 45° diagonal suggests that the model was overconfident: the low risks were underestimated and the high risks were overestimated. The Brier score (Brier 1950) gives a measure of the accuracy (reliability + resolution) of the forecast. The Brier skill score gives a measure of the accuracy of the forecast relative to a reference forecast, which in the present case is the climatology. For the reliability diagram shown in Fig. 7a, the Brier skill score was positive (0.04), indicating that the model was more skillful than climatology.

The distribution of model probabilities for the time
Fig. 6. Map of ROC scores of the probability that 2-m temperature is in the upper tercile. Only the scores over land points are shown. The terciles have been defined from the model climatology. The verification period is Mar 2002–Dec 2003. Shaded areas indicate areas where the ROC scores are higher than 0.5 (better than climatology).
range 12–18 days had a maximum close to the climatological probability and decreased toward high or low probabilities (see the circles in Fig. 7a). This is in contrast to the forecasts of the previous week (days 5–11) where the model often predicted either high or low probabilities of an event (Fig. 7b). Thus, the days 12–18 forecasts had poor resolution in comparison with medium-range weather forecasts.

The results presented above suggest that the model performed generally better than climatology for days 12–18. However, for that time range, comparing the performance of the forecast with persistence of medium-range forecasts is a much tougher test, as discussed in the introduction. Therefore, in the rest of this section, the scores of the monthly forecast over the period days 12–18 are compared to the persistence of the probabilities of the previous weekly mean anomalies (days 5–11). It is also possible to persist the ensemble mean of days 5–11 anomalies instead of persisting the probabilities. However, the ROC scores obtained by persisting probabilities are higher than those obtained by persisting the ensemble mean (not shown), suggesting that comparing the monthly forecasts with the persistence of the probabilities from the previous week is a tougher test than comparing it with the persistence of the ensemble mean of the previous week.

Figure 8a shows the ROC diagrams of the probability that the 2-m temperature is in the upper tercile. The figure shows a comparison between the monthly forecasting system for days 12–18 (black line) and persisting the 2-m temperature anomalies of the previous week (gray line). Figure 8a suggests that the monthly forecasting system performed better than persistence of the previous week. The ROC score was higher (0.67 instead of 0.62) using the monthly forecasting system than when using persistence. A scatterplot diagram of each individual case (Fig. 9a) indicates that the monthly forecasting system outperformed the persistence of the probabilities from the previous week in 70% of cases. A WMW test suggested that the scores of the monthly forecasting system were significantly higher than those obtained with persistence with a level of confidence larger than 95%. Therefore, the difference in ROC score presented in Fig. 8a was likely to be significant. A scatterplot diagram of Brier skill score (Fig. 9b) indicated an even larger difference between monthly forecast and persistence. Persistence outperformed the monthly forecast in only four cases in 45. However, the comparison of Brier skill score may not be fair for persistence, since the probabilities of the medium range were much sharper than those of days 12–18, and the Brier skill score has a tendency to penalize forecasts that are overconfident.

To address the potential benefit of the monthly forecasting system for different users, value diagnostics have been calculated, derived from a simple cost–loss model of economic decision making (Murphy 1977; Richardson 2000). In this model, a decision maker has two alternatives, to take action or to do nothing, the choice of which depends on his or her belief a given weather event will occur or not. Taking no action incurs a cost $C$ irrespective of the outcome. If the event does occur and
no action has been taken then the decision maker incurs a loss $L$. It is convenient to consider the expense of the various courses of action in term of the “cost–loss” ratio $C/L$. In this model (see Richardson 2000 for more details), the expected mean expense (ME) for the forecast system is given by the following equation:

$$
\text{ME} = \text{FAR} \frac{C}{L} (1 - \bar{\sigma}) - \text{HR} \sigma \left( 1 - \frac{C}{L} \right) + \bar{\sigma},
$$

where $\bar{\sigma}$ is the observed relative frequency, HR is the hit rate, and FAR is the false alarm rate of the probabilistic event considered in the monthly forecasting system.

The value $V$ of the forecast is defined as the savings made by using the monthly forecasting system as a fraction of the potential savings that would be achieved with perfect knowledge of the future; $V = 0$ indicates that the forecast has no more value than climatology.
Figure 8b displays the potential economic value diagram obtained with the monthly forecasting system (black curve) for the period days 12–18, and the potential economic value diagram obtained with persisting the probabilities from the previous week (gray curve). The event scored is still the probability that 2-m temperature is in the upper tercile. All land points over the Northern Hemisphere extratropics and the 45 cases were taken into account. Figure 8b confirms that the monthly forecasting system had some value for a large range of cost-loss ratios and that it had more value than persisting the probabilities from the previous week.

Previous figures in this section have shown scores for one single event, the probability of 2-m temperature in the upper tercile. However, the results did not depend strongly on the threshold. For instance, scoring the probability of 2-m temperature weekly anomaly larger than 0, 1, or 2 K (not shown) suggested the same conclusion: the monthly forecasting system had some value for days 12–18, and it outperformed persistence of days 5–11 probabilities. This conclusion was also valid with other variables. For precipitation (not shown), the ROC scores were higher for the monthly forecasting system than for persistence, although the ROC scores were much lower for precipitation than for 2-m temperature. The monthly forecasts of precipitation were more reliable than persistence, with a positive Brier skill score, suggesting that the model performed better than climatology. The difference in scores between monthly forecasts and persistence was even larger with mean sea level pressure than with 2-m temperature. Scatterplot diagrams indicated that the difference of scores between monthly forecasts of mean sea level pressure and the persistence of the probabilities from the previous weeks (days 5–11) was significant, with a level of confidence larger than 95%. Persistence outperformed the monthly forecasting system only on a few occasions. All these results suggested that overall, the monthly forecasting system produced useful forecasts for days 12–18.

However, the performance varied from one region to another. The model was particularly skillful over North America (ROC score of 0.7, and the ROC score of persistence was 0.6). Over the Southern Hemisphere extratropics (south of 20°S) the scores were about the same as over the Northern Hemisphere extratropics, suggesting that the monthly forecasting system produced useful forecasts for days 12–18. Over Europe, the potential economic value was much higher.

\[ V = \frac{\text{ME(climate)} - \text{ME(forecast)}}{\text{ME(climate)} - \text{ME(perfect)}}. \]
less than over the other regions, and the scores based on all the land points over Europe were not significantly better than those based on persistence of the previous week anomalies.

c. Skill of the system for the period days 19–32

The different fields (2-m temperature, precipitation, mean sea level pressure) were averaged over the days 19–32 (the two last weeks of the monthly forecast). Brier skill scores, ROC scores, and potential economic values were computed over land grid points and over the 45 cases. The scores were also compared to the scores obtained by persisting the probabilities of the two previous weeks of the forecast (days 5–18) so that the period persisted had the same length as the forecasting period. This is referred to as persistence.

The point map of ROC scores for the probability that the 2-m temperature anomalies were in the upper tercile (Fig. 10) suggests that over the vast majority of land points, the ROC score exceeded 0.5. This suggests that the model performed generally better than climatology. However, the scores were generally much lower than those obtained over the period days 12–18 (Fig. 6). The ROC score calculated over all the grid points in the Northern Hemisphere extratropics was slightly better than persistence (Fig. 11a). This difference seemed to be significant, with a level of confidence larger than 90% according to the WMW test applied to the ensemble of 45 cases. The model outperformed persistence in 34 of 45 cases (Fig. 12). This suggests that over the whole Northern Hemisphere extratropics, the model displayed moderate skill in predicting the probability that 2-m temperature was in the upper tercile, and it performed better than persistence and climatology.

There was some strong regional variability, and Europe was a particularly difficult region, with an ROC score of about 0.5 (Fig. 11b). On the other hand, the model displayed some skill over North America (Fig. 11c). Over North America, the model performed significantly better than climatology and persistence. Over the Tropics and Southern Hemisphere extratropics, the conclusions were about the same as for the Northern Hemisphere extratropics.

After about 20 days, the value of the forecast was very dependent on the threshold of the event. For small thresholds, such as the probability that the 2-m temperature anomaly was larger than 0, the potential economic value was quite low, and the model did not perform significantly better than persistence. However, for higher thresholds—for example, the probability that the 2-m temperature anomaly was larger than 2 K—the model displayed some value for lower cost–loss ratios, and more value than that of persistence (not shown). This suggests that even at this time range, the monthly forecasting system could still be useful. For surface temperature, precipitation, and mean sea level pressure, the conclusions were about the same as for 2-m temperature.

4. Verification of the hindcast

In addition to the real-time forecast, a 5-member hindcast was produced every two weeks in order to evaluate the model drift and calibrate the real-time forecast (see section 2). After more than one year of operations, this represents more than 500 cases spanning over 12 years. The hindcast had the inconvenience of having a small ensemble size (5), and also was based on past years where the initial conditions were probably not as good quality as the initial conditions for the real-time forecast that spans over two recent years. Therefore the scores obtained with the hindcast were likely to be lower than those obtained with the real-time forecast and could be considered to some extent as a lower bound of the skill of the real-time forecast. However, this data had a much larger number of cases (540 instead of 45) and spanned over 12 years instead of less than two years. This allowed us to investigate the seasonality of the skill of the monthly forecasting system and its interannual variability.

a. Seasonality

Table 1 displays the ROC scores for each season [December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON)] for the probability that 2-m temperature was in the upper tercile in the Northern Hemisphere extratropics. Results with other events suggested that the seasonality did not depend on the event considered. There was a strong and significant seasonality in the probabilistic scores for the period days 12–18. Over the Northern Hemisphere extratropics, the ROC score was the highest in winter. Summer was a particularly difficult season, with an ROC score of only 0.59 compared to 0.67 in winter. This result was consistent with the seasonality of the scores in the medium-range weather forecast. Winter was not the most persistent season (Table 1), and winter was the season with the strongest difference between the scores of the monthly forecasting system and the scores of persistence, making it the season where the monthly forecasting system was likely to be the most useful. In spring, the scores of the monthly forecasting system were significantly higher than the scores of persistence. In summer and fall, the scores of the monthly forecasts were just slightly higher than persistence. For the southern extratropics (not shown), results were about the same, with the model displaying the strongest skill during the Southern Hemisphere winter (JJA).

During the two last weeks of the forecast, the scores displayed less seasonality than for days 12–18 (Table 2). According to Table 2, winter still appeared to have the highest scores, but the difference with the other seasons was small. The difference with persistence was also strongest in winter. This was also true for the Southern Hemisphere extratropics (not shown).
FIG. 10. Same as Fig. 6, but for 2-m temperature averaged over the period days 19–32.
Fig. 11. ROC diagrams of the probability that 2-m temperature averaged over the period days 19–32 is in the upper tercile over (a) the Northern Hemisphere extratropics, (b) Europe, and (c) North America. The black curve represents the diagram obtained with the monthly forecasting system, and the gray curve the diagrams obtained by persisting the probabilities from the previous 2 weeks (days 5–18).
b. Interannual variability

The 45 cases used to evaluate the skill of the real-time forecasts in the previous section were taken over a period of less than 2 yr. It is possible that this period was not representative of the long-term climate, particularly if the scores of the monthly forecasting system were sensitive to low-frequency signals such as the interannual variability of SSTs. In order to answer this question, the scores of the hindcasts have been computed over the period 1990–2001. The main results obtained with the real-time forecasts (section 3) were confirmed with this study: the model displayed moderate skill for days 12–18, which was significantly higher than the skill obtained by persisting the probabilities of the previous week and the skill of climatology. The conclusion was also the same for the period days 19–32, where the skill of the model was very low, but still performed slightly better than climatology and persistence over some regions. In addition, the scores were also computed for each individual year. Results suggested that the scores displayed some interannual variability, and the scores obtained with the period 2002 and 2003 (covering the period of the real-time forecast discussed in section 3) were within the ensemble defined by the other years. This suggested that the results displayed in the previous section were quite representative of the general skill of the ECWMF monthly forecasting system.

The ROC scores of the probability that 2-m temperature is in the upper tercile were computed for each DJF from 1991 to 2002. They displayed some strong interannual variability. For instance, over North America, the ROC score for the period days 12–18 oscillated between 0.57 (DJF 1998) and 0.74 (DJF 2001). Each winter represents only 10 cases. Therefore part of the variability could be due to a sampling issue. However, the scores based on 5 of 10 cases for each winter displayed...
played similar interannual variability, suggesting that the interannual variability in the ROC score was robust. Over the Tropics, the interannual variability of the ROC scores was strongly correlated with ENSO variability. Scores were higher with higher SST anomalies. This also applied to persistence. Over North America, there was a weak but significant negative correlation ($-0.5$) between the ROC scores of 2-m temperature and the Pacific–North American (PNA) index derived from the formula in Wallace and Gutzler (1981) \[ \text{PNA} = 0.25[Z(20^\circ\text{N}, 160^\circ\text{W}) - Z(45^\circ\text{N}, 165^\circ\text{W}) + Z(55^\circ\text{N}, 115^\circ\text{W}) - Z(30^\circ\text{N}, 85^\circ\text{W})], \]

where $Z$ values are standardized 500-hPa geopotential heights from the ECMWF operational analysis and reanalysis. The lowest scores were obtained during the strong El Niño event of 1997/98. Figure 13a displays the correlation between the interannual variability of the PNA index averaged over DJF from 1991 to 2002 and the interannual variability of the ROC scores of the monthly forecasting hindcasts averaged over the same period.
ability of the ROC score of the probability that 2-m temperature is in the upper tercile over each grid point. Over large areas, the correlation was below −0.4. Over Europe, it was the North Atlantic Oscillation (NAO) index that seemed to have some impacts on the ROC scores of 2-m temperature. In the present study, the NAO index is defined as the difference between the sea level pressure over the Azores and the sea level pressure over Iceland from the NCEP–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996). The index has been averaged over December, January, and February. Figure 13b displays the correlation between the interannual variability of the NAO index averaged over DJF from 1991 to 2002 and the interannual variability of the ROC score of the probability that 2-m temperature is in the upper tercile over each grid point. The correlations in Fig. 13b are not very strong, but they seem consistent with Jung et al. (2003) and Ferranti et al. (2002). According to Fig. 13b, a positive NAO index was conductive to a reduction of the scores over northern Europe and larger scores over the Mediterranean region. This could be explained by the fact that during a positive phase of the NAO, the number of storms increased over northern Europe and decreased over southern Europe (Jung et al. 2003).

5. Conclusions and discussion

A monthly forecasting system was set up at ECMWF. This system, based on coupled ocean–atmosphere integrations, has been run routinely since March 2002. Real-time forecasts are presently produced every 2 weeks, but the frequency will become weekly by mid-2004 when the system becomes fully operational. During the 10 first days of the forecast, the monthly forecasting system produced forecasts that were close to those obtained with the operational Ensemble Prediction System. At that time range, the ocean–atmosphere coupling did not seem to significantly affect the scores.

Over days 12–18, the monthly forecasting system produced forecasts that were generally better than climatology or persistence of the probabilities from the previous week. Therefore, the monthly forecasting system is potentially useful for forecasts at this time range. Summer seems to be a particularly difficult season as in the medium range. The monthly forecasting system for days 12–18 beat persistence of days 5–11 in most cases, when using either the Brier skill score or the ROC score. During the period of verification, the model was particularly skillful over North America, central Asia, and the Southern Hemisphere extratropics. Europe looks to be a particularly difficult region at this time range.

During the two following weeks (from days 19 to 32), the coupled model performed generally better than persistence and climatology. At this time range, the model’s skill increased with higher thresholds. The model displayed some skill over North America and the Southern Hemisphere extratropics. However, the model displayed little skill over Europe.

Although the scores obtained after 10 days of forecasts were not that high, particularly when compared with the scores obtained in the medium range, the model seems to produce useful forecasts for days 12–18. This is quite an encouraging result since the previous attempt at monthly forecasting with the ECMWF system failed to produce forecasts after 10 days that were significantly better than persisting medium-range forecasts. This suggests that constant improvements in operational medium-range forecasting in the last decade may have had a beneficial impact on the extended range, making monthly forecasting potentially useful. Based on these results, the monthly forecasting system will become operational at ECMWF in 2004.

The atmospheric component of the monthly forecasting system uses the same physics as the operational medium-range forecasting system. Since the start of the monthly forecasting experiment, there have been several changes in the model physics. A new version of the atmospheric model is introduced only if it does not degrade the scores of the medium-range forecasts. A comparison of the hindcast runs produced with these different versions of the atmospheric model suggests that these changes in the model physics had beneficial impacts on the scores after day 10. Most of the results presented in this paper were produced with IFS cycle 25r1. A comparison of the hindcast produced with IFS cycle 25r1 with the hindcasts produced with a more recent version of the model (IFS cycle 26r3) in winter suggests that IFS cycle 26r3 produced significantly better forecasts over the Northern Hemisphere than IFS cycle 25r1. Scatterplot diagrams suggest that this difference was significant, with a level of confidence larger than 99%. It is of course not guaranteed that any change in the operational model physics will improve the extended forecasts, but so far this has been the case, with a clear improvement in the monthly forecasts since the start of this experiment.

There is also some potential for improvement of the monthly forecasting by using retrospective forecasts to improve extended-range forecasts. In the present configuration of the monthly forecasting system, the forecasts were just “bias corrected” by simply removing the model bias from the real-time forecast. This was a crude way of correcting model errors that does not take account of flow-dependent errors, for instance, or biases in spread. More sophisticated ways of calibrating the forecasts (see, e.g., Hamill et al. 2004) may lead to some significant improvements in the quality of forecasts, although such methods would need a much longer hindcast than is presently done. Another way of correcting model error is the multimodel approach (Krishnamurti et al. 1999). A multimodel operational seasonal forecasting system is currently being set up at ECMWF, and this approach could be applied to the monthly time scale.

This paper did not address the issue of how the model
compares to other forecasting systems, and in particular to the linear inverse model (LIM) described in Newman et al. (2003). Further plans include comparing the skill of the ECMWF monthly forecasting system with the skill of LIM, which also displays some skill in week 2 (days 8–14) and week 3 (days 15–21) (Newman et al. 2003).

A key source of predictability in the extended range is the Madden–Julian oscillation. It is very important for a monthly forecasting system to successfully predict the evolution of the MJO and its impact on the large-scale circulation. This will be discussed in a separate paper.

Acknowledgments. The author would like to thank the rest of the seasonal forecasting group for their help in setting up the system and for helpful discussions. The author would also like to thank Rob Hine, who has helped to improve the quality of the figures, and the two reviewers whose comments proved invaluable in improving the presentation of the material.

REFERENCES


