Medium-Range, Monthly, and Seasonal Prediction for Europe and the Use of Forecast Information

MARK J. RODWELL AND FRANCISCO J. DOBLAS-REYES

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

(Manuscript received 15 November 2004, in final form 26 September 2005)

ABSTRACT

Operational probabilistic (ensemble) forecasts made at ECMWF during the European summer heat wave of 2003 indicate significant skill on medium (3–10 day) and monthly (10–30 day) time scales. A more general “unified” analysis of many medium-range, monthly, and seasonal forecasts confirms a high degree of probabilistic forecast skill for European temperatures over the first month. The unified analysis also identifies seasonal predictability for Europe, which is not yet realized in seasonal forecasts. Interestingly, the initial atmospheric state appears to be important even for month 2 of a coupled forecast.

Seasonal coupled model forecasts capture the general level of observed European deterministic predictability associated with the persistence of anomalies. A review is made of the possibilities to improve seasonal forecasts. This includes multimodel and probabilistic techniques and the potential for “windows of opportunity” where better representation of the effects of boundary conditions (e.g., sea surface temperature and soil moisture) may improve forecasts. “Perfect coupled model” potential predictability estimates are sensitive to the coupled model used and so it is not yet possible to estimate ultimate levels of seasonal predictability.

The impact of forecast information on different users with different mitigation strategies (i.e., ways of coping with a weather or climate event) is investigated. The importance of using forecast information to reduce volatility as well as reducing the expected expense is highlighted. The possibility that weather forecasts can affect the cost of mitigating actions is considered. The simplified analysis leads to different conclusions about the usefulness of forecasts that could guide decisions about the development of “end-to-end” (forecast-to-user decision) systems.

1. Introduction

During the summer of 2003, Europe experienced a strong heat wave. June–August (JJA) temperature anomalies peaked at over 4 K (and five standard deviations) above the 30-yr mean over France and Switzerland (Schär et al. 2004). European rainfall was also reduced, particularly over southern France. The heat wave was apparently responsible for (or at least accelerated) the deaths of nearly 15 000 people in France alone, caused billions of euros in damage to crops, and had detrimental impacts on metropolitan pollution levels and alpine glaciers. Here, we will examine this particular case study because it is of interest in itself and because it provides a means of motivating the general topic of predictability of European climate. For example, we will use this case study to illustrate what we mean by “predictability” and to highlight some salient questions that need answering. Figure 1 shows (dashed curve) the daily climatological average European 2-m temperatures ($T_{2m}$) based on the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) record (Uppala et al. 2005). The much higher daily observed values for summer 2003 are shown by the dotted curve. Horizontal dashed and dotted lines show the corresponding weekly mean and monthly mean values. Important questions are how well was this 2003 event “predicted,” and what impact did the predictions have on decision making?

The solid black curve shows a single high-resolution forecast, traditionally known as the “deterministic forecast.” It follows quite closely the observed rise in European temperatures during the first 8 days of June. Although this high-resolution forecast was quite accurate over this medium range, it is only the ensemble
forecast (i.e., a set of equally likely forecasts, each identical except for small perturbations to the initial conditions and tendencies that reflect chaotic uncertainties) that can provide a measure of the likelihood that temperatures will rise. Since all ensemble members (gray curves) showed a temperature on 8 June \( T_{\text{8June}} \) greater than the climatology \( T_{\text{clim}} \), where the same date is implied) a prediction could be made that there is almost 100% probability that \( T_{\text{8June}} > T_{\text{clim}} \). We say “almost” because there is an assumption that the model represents faithfully the dynamics and physics of the real system and that the ensemble (with 51 members) captures the full range of chaotic uncertainty. Note that for 75% of the ensemble members \( T_{\text{8June}} \) is more than 3 K above the climatology. Hence, if we are happy to sacrifice 100% determinism, we could also have forecast the “event” \( T_{\text{8June}} > T_{\text{clim}} + 3 \) K with a probability of 75%: \( P(T_{\text{8June}} > T_{\text{clim}} + 3 \) K) = 0.75. This event also verified in the observations. More generally, for a “reliable” forecast system, an event with 75% probability of occurring would verify 75% of the time (note that “reliability,” which has a mathematical definition, is necessary but not sufficient for a good probabilistic forecasting system). For this heat wave, Fig. 1 suggests that forecast skill extends over the entire first month of the forecast when weekly averages are used (these weekly averages come from the ECMWF “monthly” coupled ocean–atmosphere ensemble forecasting system; see also Vitart 2005). For example, \( P(T_{\text{9-15June}} > T_{\text{clim}}) = 0.98 \), \( P(T_{\text{9-15June}} > T_{\text{clim}} + 1 \) K) = 0.90, \( P(T_{\text{16-22June}} > T_{\text{clim}}) = 0.80 \), and \( P(T_{\text{23-29June}} > T_{\text{clim}}) = 0.75 \) with all of these events verifying in the observations. Forecasts started from a range of days throughout JJA 2003 show similar levels of apparent skill. One may ask whether these levels of apparent skill are specific to this extreme heat wave or whether they exist more generally. (We will show in section 6 that general skill does exist throughout the first month, although perhaps not with the levels apparent for this particular heat wave.) The forecast appears to be less accurate in the second and third months of the heat wave when monthly means are used. Although the extreme nature of the heat wave means that this is not necessarily an indication of a general lack of skill, we will highlight the potential for future forecast improvements at these ranges. (These monthly averages come from the ECMWF “seasonal” coupled ocean–atmosphere ensemble forecasting system.)

At forecast ranges beyond a few days, predictability of the timing and existence of individual atmospheric synoptic systems becomes less, and predictability of the mean weather (if there is any predictability) must come increasingly from “boundary forcing.” For monthly and seasonal forecasts, although strictly interactive, sea surface temperature (SST) and soil moisture for example may provide a boundary forcing that allows some atmospheric predictability. At decadal time scales, it is possible that atmospheric predictability may arise from predictable variations in oceanic circulation or from coupled modes of “ocean–atmosphere” variability or from foreseeable changes in atmospheric composition (i.e., anthropogenic forcing).

In section 2, we undertake a systematic “unified” analysis of the skill in predicting European temperatures of the current operational (atmospheric) medium-range and (coupled ocean–atmosphere) monthly and seasonal ensemble forecast systems at ECMWF. The unified approach is fruitful in identifying future sources of forecast skill. In section 3, we concentrate on the seasonal time scale and highlight observational evidence for predictability and discuss the role of boundary forcing in this predictability. We look at the multimodel approach to improving seasonal forecast skill and attempt to quantify ultimate levels of seasonal predictability. We also review recent work aimed at identifying, and validating in models, some of the important mechanisms involved in seasonal predictability. Assuming that there is climate predictability at some time.
scale, an essential question to ask is what benefit forecasts at these time scales can have on society. If a forecast, however skillful, has no impact on decision making, one can argue that it is pointless to make it. The “value” of such forecasts to a particular user depends on their vulnerability to weather or climate anomalies and on what actions they can take to mitigate against any loss. In section 4, we investigate the value of forecast information for a range of mitigation strategies. In particular, we investigate the possibility that the cost of a mitigating action may be dependent on the forecast probability itself. In section 5, conclusions and a discussion of key issues are presented.

2. Unified analysis of European skill in the operational ECMWF ensemble prediction systems

Here, a systematic analysis of European probabilistic forecast skill over the first few weeks is made. A unified (or “seamless”) approach is taken by jointly analyzing the three forecast systems. The approach allows us to assess predictability on a large range of temporal scales and to discuss differences between the forecast systems. We concentrate on Europe and the conclusions drawn will be different for other regions of the globe. One aim is to determine if the levels of skill apparent in the forecast of the European 2003 summer heat wave described in section 1 are specific to that extreme event or are typical of predictability in general at these ranges. Another aim is to investigate the dependence of probabilistic forecast skill on forecast lead time and on the period over which a forecast is averaged [see Kharin and Zwiers (2001) for an analysis of nonprobabilistic scores]. For example, with reference to Fig. 1, one may ask whether a week is the best period of time over which to average a forecast with a lead time of 5 days?

A longer averaging period could improve skill if there was poor timing of otherwise predictable synoptic systems. A longer averaging period may also benefit the forecast by reducing the time mean of the unpredictable variability relative to a more slowly varying (and thus more constant) boundary-forced predictable signal. On the other hand, a longer averaging period would be likely to reduce the beneficial impact of the initial atmospheric conditions. The optimal averaging period for a given forecast lead time may therefore depend on a delicate balance between these competing effects.

a. Methodology

To jointly analyze medium-range, monthly, and seasonal time scales, a diagnostic is required that, while being of relevance to European users, is also comparatively represented in atmospheric high-resolution 10-day forecasts and coupled lower-resolution monthly and seasonal forecasts. Here 850-hPa temperature \( T_{850} \), averaged over the same European box as used in Fig. 1 (5°W–25°E, 35°–55°N) is chosen because it avoids land surface issues associated with differing orographies and land–sea masks used in the forecast systems described here. Data come from the three operational ensemble prediction systems at ECMWF together with analyses. More details of the model and analysis data are presented in the first five columns of Table 1.

It is possible that predictability (however defined) of \( T_{850} \) varies over the annual cycle (see later) but in order to get a sufficiently large sample we analyze all forecasts together, regardless of start date within the year. With forecast system development in mind and to make the unified analysis as consistent as possible, the model data is not “bias corrected” here. Bias correction improves the plausibility of the forecast products and is performed prior to delivery to users of the monthly and seasonal forecasts. It involves making ensembles of “hindcasts” from historical start dates with the present (most up-to-date) forecast system and calibrating the forecast with the hindcast biases.

The “event” chosen to be assessed is “European \( T_{850} > \) normal” (with “normal” defined as the climatological mean of the variable at the same time of year as the forecast). This event could be of interest to, for example, arable farmers who would like a forecast of the growth potential for the coming month and energy traders interested in future heating oil demand (\( T_{850} \) is sufficiently related to surface temperature for this theoretical study).

Here the Brier score, \( b \) (see, e.g., Palmer et al. 2000), is used to quantify probabilistic forecast skill of the event. The Brier score measures the mean squared difference between the forecast probability of the event occurring \( p_i \) at time \( t_i \) and the actual outcome \( v_i (=1 \) or \( 0 \) depending on whether the event happens or not, respectively):

\[
b = \frac{1}{m} \sum_{i=1}^{m} (p_i - v_i)^2. \tag{1}
\]

where the sum is over all forecast start dates \( i = 1, \ldots, m \) \( (m = 875, 110, \) and 44 for the medium-range, monthly, and seasonal forecast systems, respectively). The Brier score is analogous to the mean-square error used for assessing deterministic forecasts: a smaller Brier score implies more skill; a perfect forecast would have \( b = 0 \).

It is also possible to define a climatological Brier score \( b^{\text{clim}} \) for a forecast based on the climatological
<table>
<thead>
<tr>
<th>Product</th>
<th>ECMWF analysis</th>
<th>ECMWF operational ensemble prediction system</th>
<th>DEMETER seasonal forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ERA-40</td>
<td>Operational analysis</td>
<td>Medium range</td>
</tr>
<tr>
<td>Ocean model</td>
<td></td>
<td>IFS® 24R3–28R1</td>
<td></td>
</tr>
<tr>
<td>Total start dates</td>
<td></td>
<td>875</td>
<td>110</td>
</tr>
<tr>
<td>Forecast range</td>
<td></td>
<td>Days 1–10</td>
<td>Days 1–32</td>
</tr>
<tr>
<td>Bias correction</td>
<td></td>
<td>No</td>
<td>Five-member hindcasts from the same day of the year for the previous 12 yr.</td>
</tr>
<tr>
<td>Horizontal</td>
<td></td>
<td>T₁,159</td>
<td>T₁,159</td>
</tr>
<tr>
<td>resolution</td>
<td></td>
<td>(−1.125°)</td>
<td>(−0.351°)</td>
</tr>
<tr>
<td>Vertical levels</td>
<td></td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>Ensemble size</td>
<td></td>
<td>(1)</td>
<td>(1)</td>
</tr>
<tr>
<td>Initial perturbations</td>
<td></td>
<td>—</td>
<td>Atmosphere®</td>
</tr>
<tr>
<td>Stochastic physics</td>
<td></td>
<td>—</td>
<td>Yes®</td>
</tr>
</tbody>
</table>

---


*b* IFS = Integrated Forecasting System (Gregory et al. 2000).

*c* ARPEGE = Action de Recherche Petit Echelle Grand Echelle (Déqué 2001).

*d* Pope et al. (2000).

*e* HOPE = Hamburg Ocean Primitive Equation model (Wolff et al. 1997).

*f* OPA = Ocean Parallélisé model (Madec et al. 1997).

*g* GloSea = Global Seasonal model (Gordon et al. 2000).


*i* Buizza et al. (1999).
probability by replacing the model forecast probability with the climatological probability, \( p_{\text{clim}} = (1/M) \sum_{i=1}^{M} v_i \), where the sum is over the M dates included in the climatology (\( M = 16436 \) for the full ERA-40 used here):

\[
\begin{align*}
\lambda_{\text{clim}} & = \frac{1}{M} \sum_{i=1}^{M} (p_{\text{clim}} - v_i)^2 \\
& \approx p_{\text{clim}}^2 (1 - p_{\text{clim}}) \text{ for large } M. \end{align*}
\]

(2)

Assuming that the median observed \( T_{850} \) is close to the mean observed \( T_{850p} \), the climatological probability of this event occurring will be \( p_{\text{clim}} \approx 0.5 \) and thus \( \lambda_{\text{clim}} \approx 0.25 \). This is found to be a very good approximation for the event European \( T_{850} \) > normal. Hence a model Brier score less than 0.25 can be considered to represent a skillful forecasting system. Note that more elaborate (and more skillful) forecast methods can be used to judge the model against. These may involve a mix of climatology and the persistence of anomalies (used later in this paper to assess seasonal forecast skill).

Each ensemble forecast is initiated from the 0000 UTC analysis and we assess the skill of the forecast at 1200 UTC on the start date and for subsequent days. For each ensemble forecast system in turn (medium range, monthly, and seasonal), all possible combinations of forecast lead time (LT) and averaging period (AP) are considered. For simplicity, we write LT = 1, 2, 3, . . . for lead times of 12, 36, 60, . . . hours. The average, for example, over the three lead times 36, 60, and 84 h is then denoted as LT = 2, AP = 3. We can do this for LT \( \geq 1 \), AP \( \geq 1 \), and LT + AP - 1 \( \leq R \) where R is the range of the forecast (10, 32, and 184 days for the three systems, respectively). Hence, a function \( f_n(LT, AP) \) will be defined for (LT, AP) within the triangle bounded by these three inequalities. (Note that the event “European \( T_{850} \) > normal” was chosen because with any other threshold, normal + x, for example, \( x \) would need to be a function of AP and would lead to unnecessary complication of the analysis.)

If one assumes that the highest-resolution forecast system produces the best European \( T_{850} \) forecast (it generally does) then one should use the medium-range system for a forecast with LT + AP - 1 \( \leq 10 \), the monthly system for a forecast with 10 \( < \) LT + AP - 1 \( \leq 32 \) and the seasonal system for 32 \( < \) LT + AP - 1 \( < 184 \). Hence in Fig. 2 we show “montages” of forecast diagnostics as a function of LT and AP with the medium-range forecast scores shown in the triangle LT + AP - 1 \( \leq 10 \), the monthly forecast scores shown for 10 \( < \) LT + AP - 1 \( \leq 32 \), and the seasonal forecast scores shown for 32 \( < \) LT + AP - 1 \( \leq 184 \) (with LT and AP limited to 44 days).

b. Medium-range and monthly forecast results

1) PROBABILISTIC FORECAST SKILL

Figure 2a shows the Brier score montage. Unsurprisingly, the smallest Brier scores (<0.05) occur for smallest LT and AP, indicative of the impact of initial conditions of the atmosphere on the medium-range forecast. The monthly forecast system shows similar levels of skill in the medium range (LT + AP - 1 \( \leq 10 \)) to that of the medium-range forecast (not shown). This is reflected in the fact that there is no discontinuity apparent in Fig. 2a at LT + AP - 1 = 10 (the white dashed line that divides the display of medium range from monthly forecast results).

Presently, ECMWF operationally delivered monthly forecast products focus on AP = 7 (1-week averages) and LT = 5, 12, 19. The corresponding Brier scores are 0.12, 0.20, and 0.23, respectively. The probabilities \( p(T_{850} > T_{850}^{\text{clim}}) \) from all the monthly forecasts made during the summer heat wave are shown in Table 2. Although this is quite a small set of forecasts (seven for each lead time), the Brier scores based on just these forecasts are 0.07, 0.07, and 0.08, respectively. Since these values are less than those for the systematic analysis, it is possible that this extreme event may have been particularly predictable. Some of this predictability may reflect the persistence of preexisting warm anomalies. However, several intraseasonal variations were also predicted. For example, the small rise in probabilities with lead time from 90% to 94% for the 21 May forecast (see Table 2) reflects a good simulation of the heat wave onset at a lead time of 12+ days (this is better seen in the predicted temperature anomalies than the probabilities; not shown). The particularly rapid decline in probabilities for the 18 June forecast (from 96% to 61%) reflects a good forecast, with lead time 19+ days, of the cool spell in the first week of July. The rapid decline in probabilities for the 13 August forecast (from 100% to 49%) predicted the end of the heat wave.

Figure 2a shows that, for LT + AP - 1 \( \leq 32 \), the Brier score for any given lead time generally increases monotonically with increasing averaging period [this is in general agreement with the deterministic forecast skill results of Kharin and Zwiers (2001)]. To first order therefore, the above hypothesized concept of an “optimal averaging period” does not occur, or rather it is trivially 1 day for all lead times. Nevertheless, the decrease in forecast skill with increasing averaging period is remarkably slow [e.g., the Brier score is still only 0.16 at (LT, AP) = (1, 32) (the prediction of the average temperature for the entire next month)]. Presumably,
Fig. 2. Results from a joint analysis of ECMWF medium-range, monthly, and seasonal ensemble forecasts of European average temperature at 850 hPa ($T_{850}$). Quantities are defined in the main text and plotted here as a function of forecast lead time ($LT$) and forecast averaging period ($AP$). Medium-range forecast data are plotted for $LT + AP = 1 \leq 10$, monthly forecast data for $LT + AP = 1 \leq 32$, and seasonal forecast data for $32 < LT + AP < 184$ (limited to $LT \leq 45$, $AP \leq 45$). The forecast data were not bias corrected prior to this analysis. (a) Brier scores for the event “European $T_{850}$ > normal” (“normal” is defined as the ERA-40 climatology) with contour interval CI = 0.05, (b) mean model bias with CI = 0.2 K, (c) model “activity” divided by observed activity with CI = 0.05, (d) model “spread” with CI = 0.5 K, (e) RMSE with CI = 0.5 K, and (f) spread/RMSE with CI = 0.1. The dashed white line in (a) shows the boundary between the medium-range forecast results (above) and monthly forecast results (below). The boundary between the monthly forecast results and the seasonal forecast results is evident from the 0.25 contour. The European region used is (35°–55°N, 5°W–25°E). See Table 1 for more details on the operational forecast data used.
Table 2. The operational forecast probabilities (%) that weekly averaged European 850-hPa temperatures would be above the ERA-40 climatological mean value during the summer heat wave of 2003. Also printed are the Brier scores averaged over these forecasts and the Brier scores averaged over all 110 monthly forecasts made since 27 March 2002. The forecast probability in parenthesis “(49)” is not included in the Brier score for the heat wave as the heat wave had finished by then.

<table>
<thead>
<tr>
<th>Forecast start date</th>
<th>Probability European $T_{850}\geq$normal Probability (D + 5 to D + 11)</th>
<th>Probability (D + 12 to D + 18)</th>
<th>Probability (D + 19 to D + 25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 May 2003</td>
<td>—</td>
<td>—</td>
<td>82</td>
</tr>
<tr>
<td>21 May 2003</td>
<td>90</td>
<td>94</td>
<td>73</td>
</tr>
<tr>
<td>4 June 2003</td>
<td>96</td>
<td>78</td>
<td>71</td>
</tr>
<tr>
<td>18 June 2003</td>
<td>96</td>
<td>92</td>
<td>61</td>
</tr>
<tr>
<td>2 July 2003</td>
<td>31</td>
<td>39</td>
<td>55</td>
</tr>
<tr>
<td>16 July 2003</td>
<td>98</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>30 July 2003</td>
<td>100</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>13 August 2003</td>
<td>100</td>
<td>80 (49)</td>
<td></td>
</tr>
<tr>
<td>$b_{2002-2003}$</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>$b_{2002-2005}$</td>
<td>0.12</td>
<td>0.20</td>
<td>0.23</td>
</tr>
</tbody>
</table>

where $T_{i}$ refers to the analysis temperature at the verifying time of the corresponding forecast. Figure 2b shows the montage of mean forecast biases. The medium-range and monthly forecasting systems do not show strong biases (less than 0.3 K for LT + AP $-1 \leq 19$). Simple calibration (i.e., bias correction) of the monthly forecast system based on five-member hindcasts of the 12 previous years, starting from the same day of the year, largely removes these biases but does not generally improve the Brier scores. Clearly, biases could also have a detrimental impact on the prediction of the anomalous circulation, and model improvements that reduce biases without the need for calibration are preferred.

3) Forecast error, spread, and uncertainty

We define the root-mean-square error (RMSE) of the ensemble to be

$$\text{RMSE} = \sqrt{\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (T_{ij} - T_{i}^{an})^2}.$$  \hspace{1cm} (5)

There are several ways to define the “spread” of an ensemble. One definition is the root-mean-squared difference from a control (unperturbed) forecast (Buizza 1997). Here we avoid the concept of a control forecast and define the spread of the ensemble to be the root-mean-squared difference between all possible pairs of ensemble members:

$$\text{SPREAD} = \sqrt{\frac{1}{mn(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} (T_{ij} - T_{ik})^2} = \sqrt{\frac{2}{mn(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (T_{ij} - \overline{T})^2}.$$  \hspace{1cm} (6)

As the second form in (6) demonstrates, this definition of spread is proportional to the root-mean-squared difference from the ensemble mean (i.e., the standard deviation of the ensemble members). For a “perfect” ensemble forecasting system, the observations (i.e., analysis) could be considered to be another realization of the model, and thus, for large values of $m$ (i.e., many forecast start dates) the spread should equal the RMSE. This is equivalent to saying that, when averaged over many forecasts, the mean squared difference between any two ensemble members should be equal to the mean squared difference between any one ensemble member and the analysis. Figure 2f shows that spread/RMSE is indeed approximately equal to 1 for the medium-range and monthly forecast results (it generally lies between 0.95 and 1.05). This match for all lead
times suggests that the aggregate effects of uncertainty (see below) are reasonably well captured in the ensemble system.

Uncertainty in the initial atmospheric conditions is accounted for in the medium-range and monthly ensemble systems by adding perturbations to the control initial conditions for each ensemble member. The initial perturbations are defined using singular vectors of a linear approximation of the ECMWF model (Buizza and Palmer 1995). Since positive and negative forms of each singular vector are used in the construction of the ensemble perturbations, the initial ensemble mean is equal to the analysis ($T_i = T_{an}$), and hence, from the RMSE definition ($\text{spread/RMSE} = \sqrt{2n/(n - 1)}$) at the initial time ($=1.43$ for an $n = 51$ member ensemble). It is perhaps surprising that our results show that just 12 h into the forecast spread/RMSE has approached 1 [spread/RMSE = 1.009 for (LT, AP) = (1, 1) in the medium-range forecast] for European $T_{seis}$. The adjustment to a value of 1 may not be so rapid for other fields such as 500-hPa geopotential height for example (R. Buizza 2004, personal communication).

Uncertainty in the subgrid-scale tendencies is also represented in the medium-range and monthly ensemble forecasting systems. This uncertainty arises because the large-scale (resolved) flow is not thought to be sufficient to completely constrain the subgrid-scale processes. To account for this uncertainty, a “stochastic multiplier” is presently applied to the parameterized tendencies (Buizza et al. 1999). Shutts (2005) discusses the limitations of this stochastic multiplier methodology and the potential benefits of using a “kinetic energy backscatter algorithm.” There are other potential sources of uncertainty that are not explicitly represented at present. For example, initial and stochastic uncertainty associated with land processes are neglected.

Note that spread/RMSE $\approx$ 1 is an important property of an ensemble forecasting system but it does not necessarily imply that there is no more skill to be gained from further system improvements (both spread and RMSE could be reduced simultaneously).

c. Seasonal forecast results

1) Probabilistic forecast skill

A prominent feature of Fig. 2a is the discontinuity in Brier scores at LT + AP − 1 = 32 (the interface between the monthly and seasonal forecast data). Whereas the monthly forecast has a Brier score of 0.16 at (LT, AP) = (1, 32), the seasonal forecast shows a Brier score of 0.25 at (LT, AP) = (1, 33). Comparison of both forecast systems for exactly the same LT and AP (possible for LT + AP − 1 < 32) confirms the superiority of the monthly system. Clearly the seasonal model does show better forecast skill elsewhere over the globe (such as in the Tropics; Anderson et al. 2003; Palmer et al. 2004) but this study is only concerned with Europe. This unified analysis suggests that work to reduce the skill discontinuity will result in improved seasonal forecasts for Europe, particularly for a temperature forecast averaged over, say, weeks 2–6. With reference to the other panels in Fig. 2, we use the unified approach to examine the reasons behind this discontinuity in European forecast skill.

2) Forecast biases

Figure 2b shows that the mean model bias (averaged over all forecast start dates and so, in particular, averaged over the annual cycle) also increases strongly as we move to the seasonal forecast data. One possible reason for the increased bias is that the seasonal forecasting system uses an older version of the IFS model. As one would expect, bias correction of the seasonal forecast output (using five-member ensembles of hindcasts for the years 1987–2001) prior to the analysis does reduce the forecast bias but it has little impact on the Brier score discontinuity (not shown). This is not the case for tropical atmospheric predictability as, for example, Brier scores for the Southern Oscillation index are improved with bias correction.

3) Forecast error, spread, and uncertainty

A discontinuity at LT + AP − 1 = 32 is also seen in the RMSE Fig. 2e. This is not significantly improved with bias correction. The seasonal forecast system also shows reduced spread compared to the monthly forecast system for LT + AP − 1 = 32 and this is evident as a discontinuity at LT + AP − 1 = 32 in Fig. 2d (see the kink in the contours). Reflecting both RMSE and spread deficiencies, spread/RMSE (Fig. 2f) can be much less than 1 for the seasonal forecast data [e.g., spread/RMSE = 0.78 for (LT, AP) = (1, 33)].

Another quantity of interest is the strength of transient activity. “Activity” is defined here as the root-mean-square of the difference between the forecast (or observations) and the observed climatology:

$$\text{MODEL ACTIVITY} = \sqrt{\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (T_{ij} - T_{ij}^{\text{clim}})^2},$$

(7)
where $T_{i}^{\text{clim}}$ is the climatological temperature at the same time of year as the corresponding forecast data. Figure 2c shows the model activity divided by observed activity. A value of less than 1 (more than 1) would indicate that transient activity is weaker (stronger) in the model than in reality. It can be seen that model activity is generally between 0.95 and 1.15 and therefore quite reasonable on all averaging time scales shown. Values across the boundary between the monthly and seasonal forecasting systems indicate that the seasonal system is slightly more active than the monthly system. (This is also the case, but less marked, when the monthly and seasonal forecasts are bias corrected prior to the analysis).

Increased activity is likely to have an increasing effect on RMSE and spread. Hence, the increased RMSE in the seasonal forecasts may be partly a consequence of this increased activity may explain the increased RMSE in the seasonal forecasts, but it does not explain the decreased spread. One hypothesis is that this may be due to a lack of uncertainty in the initial conditions. In fact, the operational seasonal forecasting system does not use any initial atmospheric perturbations. (The seasonal forecasts also lack stochastic physics). To test the hypothesis concerning initial atmospheric perturbations, two sets of 34 five-member seasonal coupled forecasts were made, one set without atmospheric perturbations and one set with singular vector perturbations applied. This was the only difference between each set of ensemble forecasts. The forecasts were initiated from 1 April and 1 October for the years 1987 to 2003 using the atmospheric model version 28R3. Figure 3 shows spread/RMSE results without (Fig. 3a) and with (Fig. 3b) initial atmospheric perturbations. The atmospheric perturbations clearly remove the very low values of spread/RMSE for the first two weeks of the forecasts. However, the results also confirm (what was not intuitively recognized before) that atmospheric perturbations can have an influence on the ensemble spread even for month 2 of a forecast (e.g., LT = 30, AP = 30). For LT + AP − 1 > 32 (and LT ≤ 45, AP ≤ 45), there is generally a 10%–15% increase in spread associated with the introduction of initial atmospheric perturbations. It is not clear at present whether coupling to the ocean is important for the increased spread or whether it is purely an atmospheric effect. Such initial atmospheric perturbations are likely to be included in the next seasonal forecasting system and thus representing one step toward a more seamless forecasting approach. The set of forecasts is too small to obtain estimates of Brier score differences between these two experiments.

These results suggest that a simple replacement of the older atmospheric model component of the seasonal forecast model with a newer version (and introduction of atmospheric perturbations) would improve European seasonal forecasts. While this may well be true for Europe, one needs to consider the coupled model’s ability to predict globally. In particular the representation of the El Niño–Southern Oscillation (ENSO)
is a key factor and this has prevented replacement of the atmospheric component up until now (T. Stockdale 2004, personal communication).

3. Improving seasonal predictability

This section is devoted to identifying what other improvements to a seasonal prediction system are likely to lead to improved forecasts. There is a stronger review character to this section. We will examine more closely the sources and mechanisms by which seasonal predictability may arise but we start with a discussion of some of the observational evidence for seasonal predictability.

a. Observed European predictability

Persistence of anomalies is perhaps the simplest form of prediction. Here we calculate the temporal anomaly correlation skill for the prediction of European sea-

sonal-mean $T_{2m}$ anomalies based on persistence. The data come from ERA-40 for the years 1959–2001 and anomalies are relative to the ERA-40 climatology. The predictor is the monthly mean $T_{2m}$ anomaly two months prior to the start of the season. So, for example, the October-mean anomaly is used to predict the December–February-mean anomaly. (Hereafter a season such as December–February will be denoted by the first letters of its constituent months, i.e., “DJF.”) Figure 4 shows persistence forecast temporal anomaly correlation skill for the four standard seasons. Shading indicates approximate pointwise statistical significance at the 90% level using a one-sided $t$ test. The highest levels of persistence skill for MAM (Fig. 4a) occur around the North Sea, the Baltic Sea, and along the spring snow line into central Europe. If this skill has a local origin then possible candidates for its forcing may include the influence of coastal SSTs, the albedo of snow, and the latent heat effects of melting. For JJA (Fig. 4b) the
highest persistence skill occurs over southern Europe and, if locally forced, could possibly arise from changes in the Bowen ratio associated with soil moisture anomalies in the more arid Mediterranean area. There appear to be reasonable levels of persistence skill (anomaly correlations of ~0.3) for much of western Europe in SON (Fig. 4c). It is less clear what local effects could lead to this skill, and the role of Atlantic SSTs may be important (see below). For DJF, atmospheric internal variability is strong and this may explain the reduced (even negative) skill of the persistence forecast (Fig. 4d). It is possible that some of the persistence skill seen in Fig. 4 is derived from long-term trends or low-frequency variability in the data. However, removal of linear trends in all datasets prior to analysis has very little impact on the figures produced and does not change any of the conclusions made.

Observational evidence for a role for SST in seasonal forecasting (for Europe) was examined by Czaja and Frankignoul (2002). Using a lagged maximal covariance analysis technique they found a significant influence of Atlantic SST on the North Atlantic Oscillation (NAO; Hurrell 1995) in late autumn/early winter (NDJ). (Warm subtropical SSTs and cool SSTs off the east coast of the United States coincide with a negative NAO anomaly.) Rodwell and Folland (2002) suggested that SSTs in the preceding May provided the best predictor of the subsequent DJF NAO, yielding a statistically significant correlation skill of 0.45. Rodwell and Folland hypothesized that the insulation (after May) of upper-ocean temperature anomalies below a shallow summer layer and their subsequent mixing back to the surface in the following autumn and winter (see, e.g., Namias and Born 1970; Alexander and Deser 1995) may provide the memory for the forecast. Kushnir (1994), following Bjerknes (1964), suggested that interdecadal time scale changes in SST associated with, for example, the Thermohaline Circulation (McCartney and Talley 1984) and the Gulf Stream/gyre circulation (Greatbatch et al. 1991) may have been responsible for the strong rise around 1970 in mean sea level pressure (MSLP) over the area 40°–50°N, 20°–40°W although it is unclear whether this implies any decadal predictability. Seasonal and longer time scale predictability may also come from the influence of SSTs from other regions such as the Indian Ocean (Hoerling et al. 2001), the South Atlantic (Robertson et al. 2000), and those of the El Niño–Southern Oscillation (ENSO) (e.g., Rodwell and Folland 2002). Van Oldenborgh et al. (2000), for example, report a European MAM precipitation signal that appears to be a lagged response to DJF ENSO SST anomalies.

Recent interest has also focused on the role of stratospheric “harbingers” of tropospheric weather regimes (Baldwin and Dunkerton 2001; Baldwin et al. 2003) that appear to lead to some predictability for the Arctic Oscillation and NAO.

Observational studies, such as those mentioned above, are essential if we are to validate our climate models and benchmark the skill of their forecasts. On the other hand, the shortness of the observational record and the inability to perform sensitivity studies with observational data means that model-based studies are also required to investigate the mechanisms through which the predictability arises. A discussion of mechanisms and model validation is given in section 3e. First, however, we discuss the seasonal predictability that present models display.

b. Boundary-forced potential predictability

In the real world, there is two-way ocean–atmosphere coupling at the intraseasonal time scale (Barsugli and Battisti 1998; Bretherton and Battisti 2000) and so SSTs cannot be strictly considered to be boundary conditions for the seasonal-mean atmosphere. (The use of a lag in the observational analyses above is an attempt to avoid complications associated with such two-way coupling.) Nevertheless, for atmospheric model intercomparison and for making a first estimate of “potential predictability” (the upper bound of true predictability), it has proved useful to treat the observed SSTs as boundary conditions to atmospheric model simulations. Potential predictability can be estimated by applying the analysis of variance (ANOVA) technique (e.g., Rowell 1998) to ensembles of atmospheric model simulations (which differ only in their initial conditions). Results from many models suggest that there is relatively high potential predictability of seasonal-mean atmospheric anomalies in the Tropics and subtropics but that this drops off rapidly as we move to midlatitudes (Kushnir et al. 2002). Results from atmospheric models within the EC-funded PREDICATE project suggest that less than 20% of the variance of decadally filtered MSLP, $T_{2m}$ or precipitation at each European grid point can be explained by SST forcing. Iceland may be a region of relatively high seasonal potential predictability of MSLP, and this is consistent with potential predictability studies of the NAO (e.g., Rodwell et al. 1999; Mehta et al. 2000; Doblas-Reyes et al. 2003). There are, however, large differences between potential predictability estimates from different models, even in the Tropics, and these differences need to be better understood.

Other boundary conditions may also be able to force an atmospheric response and thus augment seasonal predictability. Model sensitivity studies (Deser et al. 2004)
suggest that Arctic sea ice anomalies affect surface heat fluxes and can lead to a direct small-scale baroclinic response and a larger-scale indirect barotropic response with similarities to the NAO. Land properties such as soil moisture and snow depth also show significant persistence at the monthly to seasonal time scale and may be considered as partial boundary conditions for monthly and seasonal forecasts. Although the initial soil moisture index in ECMWF operational forecasts was dry prior to, and during, the European summer heat wave of 2003 (the European average value was around 75%), there is the possibility that the soil was even drier in reality. Ferranti and Viterbo (2006) showed that reducing soil moisture further in the initial conditions of seasonal forecasts lead to warmer (+2° to +5°C), more realistic, $T_{2m}$ throughout the season. Experiments also show improved simulation of seasonal precipitation anomalies over Eurasia when the observed (rather than climatological) soil moisture is prescribed (Dirmeyer 2000; Douville and Chauvin 2000; Douville 2002). Improvements in soil moisture observation (Houser et al. 2004; IGPO 1998) may, therefore, lead to improved seasonal temperature and precipitation forecasts. However, the use of interactive soil moisture does not appear to improve predictability over continental areas including Europe (Douville 2004) although it does lead to increased climate variability (Delworth and Manabe 1988, 1989; Koster et al. 2000; Schubert et al. 2004, based on a study of the U.S. Great Plains).

c. Multicoupled model approach to forecasting

The European Community (EC)-funded Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) project (Palmer et al. 2004) was aimed at investigating “end-to-end” seasonal predictability in coupled models and constructing a multimodel operational seasonal forecasting system. Within the DEMETER project seven quasi-independent coupled ocean–atmosphere models developed in different European scientific institutions were used in a multimodel ensemble experiment. For each standard season, a set of nine-member ensemble seasonal hindcasts were performed over the period 1980 to 2001. Each seasonal forecast was initiated from the first of the month preceding the season in question (e.g., from 1 February for MAM). For three of the models, seasonal hindcasts were made for a longer period (1959–2001). The experiment is described in detail in Palmer et al. (2004). The ocean model component of each coupled model was initialized by being forced with ERA-40 surface fluxes and a strong relaxation to observed SSTs. The atmospheric model component (and soil moisture) of each coupled model was initialized with ERA-40. The ensemble was generated using oceanic perturbations. Figure 5 shows spatial anomaly correlation coefficients (area-weighted spatial correlations between forecast anomalies and observed anomalies) for European $T_{2m}$ (Fig. 5a) and precipitation (Fig. 5b) from the ensemble means of the seven individual DEMETER models and the simple multimodel ensemble forecast. Each DEMETER model ensemble forecast has nine members. Forecast anomalies are relative to the mean over all 22×9 forecasts considered (so implicitly include a bias correction) and verifying anomalies are relative to the mean of the 22 seasons in ERA-40. For each season and model (or multimodel), the anomaly correlation is based on the 22-yr mean covariance with the observed anomalies. Note that the anomaly correlation for the multimodel MAM precipitation is zero. An anomaly correlation of $-0.28$ is statistically significant at the 90% level using a one-sided $t$ test.

![Fig. 5. Seasonal evolution of the spatial ensemble-mean anomaly correlation coefficient for (a) 2-m temperature and (b) precipitation from each of the seven individual DEMETER models (open bars) and the multimodel mean (solid bars) over Europe (35°–75°N, 12.5°W–42.5°E). The results for DJF come from hindcasts initiated on 1 November for the years 1980–2001. The same 22 yr and lead time are used for the other forecasts. Each DEMETER model ensemble forecast has nine members. Forecast anomalies are relative to the mean over all 22×9 forecasts considered (so implicitly include a bias correction) and verifying anomalies are relative to the mean of the 22 seasons in ERA-40. For each season and model (or multimodel), the anomaly correlation is based on the 22-yr mean covariance with the observed anomalies. Note that the anomaly correlation for the multimodel MAM precipitation is zero. An anomaly correlation of $-0.28$ is statistically significant at the 90% level using a one-sided $t$ test.](image URL)
within the box, but what is important here is that the multimodel mean appears to perform better than each individual model (a value of ~0.28 is significant at the 90% level, and this is achieved for the multimodel in three of the four seasons). Similar conclusions about the benefits of the multimodel can be drawn for the precipitation forecasts (Fig. 5b) although it should be noted that precipitation ensemble-mean skill is very small. Possible reasons for the apparent benefits of the multimodel approach are discussed later.

The DEMETER coupled model seasonal forecasts have a similar lead time (1 month) to that of the simple persistence forecasts (Fig. 4). Hence we can compare the seasonal persistence forecast skill with that of the three DEMETER coupled models that were run over the same set of years 1959–2001. (See the last three columns of Table 1 for further details of these three models). Figure 6 shows the temporal anomaly correlation skill of the multimodel mean, nine-member ensemble-mean for these years. It can be seen that the skill of the multimodel mean (Fig. 6) has a similar annual cycle to that of the simple persistence forecasts (Fig. 4). The three-model mean is perhaps marginally more skillful than the persistence forecast, particularly over the Atlantic. This suggests that the multimodel is capturing the predictability associated with persistence (possibly locally forced by coastal SST, snow cover, and soil moisture) that appeared to be apparent in the observations. As with the persistence forecasts, the removal of linear trends prior to the analysis has very little impact on the figures produced (not shown).

The general improvement in skill obtained by using
the multimodel is more obvious in a probabilistic setting. To demonstrate this, we calculate the Brier skill score (BSS), which relates the Brier score of the ensemble forecast, \( b \), to the Brier score based on a climatological probability forecast, \( b_{\text{clim}} \), by the equation

\[
\text{BSS} = 1 - \frac{b}{b_{\text{clim}}}. \tag{9}
\]

A value of BSS greater than 0 implies more skill than a climatological probability forecast. A value of 1 occurs if \( b = 0 \) and implies a perfect deterministic forecast.

The two events chosen to assess are “European winter \( T_{2m} \) warmer than the climatological median value” and “European winter precipitation greater than the climatological median value.” Since the median is used, \( p_{\text{clim}} = 0.5 \) and \( b_{\text{clim}} = 0.25 \).

A more complex probabilistic persistence forecast (than that used above) has been constructed to compare the models against. The persistence forecast probability, \( p_{\text{pers}}(t) \), is now a function of lead time (in months); \( p_{\text{pers}}(0) \) is set to 1 or 0 depending on whether the event occurs or not in the mean of three predictor months. Eight months later, the memory of any event is considered to be lost so that \( p_{\text{pers}}(8) = p_{\text{clim}} \). Between 0 and 8 months, \( p_{\text{pers}} \) is linearly interpolated between \( p_{\text{pers}}(0) \) and \( p_{\text{clim}} \). This more complex persistence forecast is called here the “decayed persistence forecast.” For a decayed persistence forecast of MAM based on the predictor months NDJ the probability of the event occurring is \( p_{\text{pers}}(3) = (5/8)p_{\text{pers}}(0) + (3/8)p_{\text{clim}} \). A Brier score for this persistence forecast, \( b_{\text{pers}} \), is defined as

\[
b_{\text{pers}} = \frac{1}{m} \sum_{i=1}^{m} (p_{i} - v_{i})^{2}, \tag{10}
\]

and the BSS of the persistence forecast can be calculated as

\[
\text{BSS}_{\text{pers}} = 1 - \frac{b_{\text{pers}}}{b_{\text{clim}}}. \tag{11}
\]

Figure 7 shows the BSS for the two events “European winter \( T_{2m} \) warmer than the climatological median value” (closed bars) and “European winter precipitation greater than the climatological median value” (open bars) for each individual DEMETER model (S1–S7), the simple multimodel ensemble (MM), and the decayed persistence forecast (P). We have already indicated that winter is perhaps the hardest season to predict for Europe and with this measure it is clear than none of the individual models or the decayed persistence method can beat climatology. Although modest in absolute terms, the multimodel ensemble outperforms each individual model and the persistence forecast (it has positive Brier skill scores). Part of the increased skill over the individual models can be attributed to the larger ensemble size although Palmer et al. (2004) showed that on average 54 members taken from the seven-model multimodel ensemble give better Brier skill scores than a 54-member ensemble created with just the best-performing individual model. The superiority of the simple multimodel (Hagedorn et al. 2005) over the individual models and persistence is due mainly to an increase in ensemble spread and the cancellation of model errors.

d. **Coupled model potential predictability**

Potential predictability can be estimated for each individual coupled model by taking each ensemble member in turn, assuming it represents the truth and comparing the other ensemble members with it. With this approach, one hopes to determine, with an imperfect model, an upper-bound for real predictability that would be approached as models were improved. We find that the estimates of potential predictability are highly variable between the individual DEMETER models and the differences appear to be mainly related to the choice of atmospheric model component. This suggests that we do not at present have a reliable estimate of the true potential skill of seasonal forecasts. Less models have been used to estimate potential predictability on longer (decadal) time scales and so conceivably there could be even larger uncertainties associated with these estimates. Nevertheless Collins (2002), using the Third Hadley Centre Coupled Ocean–
Atmosphere GCM (HadCM3), did find a statistically significant anomaly correlation of 0.82 for 5-yr-averaged North Atlantic SST (and 0.22 for European surface air temperature). Whether the model does capture well enough features such as the variability of the Thermohaline Circulation, its relationship with SST and the atmospheric response to SST forcing is not straightforward to validate because of the lack of observations.

e. Mechanisms and model validation

Although the North Atlantic atmosphere does appear to respond to SST, the ANOVA technique does not indicate which ocean basin is most important for the forcing. Using other techniques, Rodwell et al. (1999) suggested a role for SST anomalies in the Atlantic basin, whereas Hoerling et al. (2001) suggested an important role for low-frequency changes in tropical Pacific and Indian Ocean SST, with little role for the Atlantic. Cassou and Terray (2001) suggested roles for both ENSO and Atlantic SST although it was unclear whether these represented independent forcing mechanisms. Sutton and Hodson (2003) used an optimal detection method (Venzke et al. 1999) that can identify which regions of the ocean are most important for forcing the atmospheric model. Their SST pattern and NAO-like response have strong similarities with those of the observational lagged maximal covariance analysis results mentioned above (and thus emphasizes the role of the Atlantic). When low frequencies were analyzed separately, the optimal SST pattern was more uniform over the whole North Atlantic. If this pattern is associated with the Thermohaline Circulation (Delworth and Mann 2000) then any decadal predictability of the Thermohaline Circulation may imply predictability of North Atlantic-European climate. A high-frequency analysis emphasized a dual influence of ENSO and tropical Atlantic SST.

To investigate the apparent high sensitivity of potential predictability estimates to the choice of atmospheric model, Rodwell et al. (2004) conducted some highly controlled experiments. They used five PREDICATE atmospheric models to determine more clearly model differences and the mechanisms of the atmospheric response to imposed North Atlantic SST anomalies. They found that the magnitude of the multimodel mean response was stronger than the intermodel spread although significant model differences, even in the Tropics, were apparent. Much of the large-scale response appeared to be forced by the tropical part of the SST anomalies, possibly through the convective forcing of Rossby waves (Hoskins and Sardeshmukh 1987; Ambrizzi and Hoskins 1997) and is in agreement with the results of Terray and Cassou (2002). For the response over Europe, SST anomalies in the Caribbean region and over the extratropical North Atlantic appeared to be important. Although Europe does not stand out in ANOVA results as a region that is generally affected by SST variability, this experiment showed that there could be “windows of opportunity” when a strong (possibly predictable) European SON and DJF \( T_{2m} \) signal could arise for particular North Atlantic SST anomaly patterns. As speculated above, it is possible that the persistence skill for European temperatures in SON (Fig. 4c) could be a consequence of this forcing (see their Fig. 2a). The model response patterns were also found to be generally similar to the observational lagged maximal covariance patterns mentioned above, but somewhat weaker.

There is also a need to understand and validate surface fluxes of heat, momentum, and freshwater in coupled models. Frankignoul and Kestenare (2002) found a damping of SST anomalies in the observations, particularly in the midlatitude winter when strong climatological wind speeds enhance the influence of anomalous SST on surface turbulent heat flux and lead to values in excess of \(-40 \text{ W m}^{-2} \text{ K}^{-1}\). This is in agreement with AGCM results of Peng et al. (1997) and Rodwell et al. (1999) but in disagreement with the positive feedbacks in the hypothesized decadal oscillation of Latif and Barnett (1994)]. Frankignoul et al. (2004) found that this midlatitude damping was substantially underestimated in several coupled models. In particular, this could reduce the amplitude of an atmospheric response. These and other results suggest that there is still substantial scope for improvement of models and, possibly, of predictability estimates themselves.

4. Mitigating actions

Arguably a forecast, however accurate, is only useful if it can have a positive impact on decision making. It is clear that medium-range weather forecasts and anthropogenic climate change forecasts do have such an impact. For example, ECMWF medium-range forecasts during the European summer heat wave of 2003 were used to guide decisions on nuclear power production from river-cooled power plants (R. Mureau 2004; personal communication). However, the levels of skill associated with monthly, seasonal, and decadal forecasts for Europe may be such that to demonstrate usefulness, the whole “end-to-end” forecast-to-user-decision process needs to be optimized for each individual user (Pielke and Carbone 2002). For example, for the agricultural industry, forecasts of “growing days” (a function of temperature, soil moisture, and insolation) may...
have skill in predicting growth while forecasts of temperature alone may not. For a particular user the forecast \(P(T_{\text{JJA}} > T_{\text{clim}} + 3 \, \text{K}) = 0.75\) may be of more use than \(P(T_{\text{JJA}} > T_{\text{clim}}) = 1.00\) despite being less deterministic. This may be because the user can readily cope with a mildly warm summer but not with an excessively warm one. If the “event” \(T_{\text{JJA}} > T_{\text{clim}} + 3 \, \text{K}\) were to occur, then the user may make a substantial loss unless they take action beforehand. This action (termed here “mitigating action”) will itself have a cost associated with it. The user could optimize a critical probability \(p_{\text{crit}}^t\), which, if exceeded by the dynamical forecast probability, should lead to the mitigating action being taken. This critical probability will depend on the loss, the cost, and on the ability of the forecasting system to predict the event. This approach was discussed by Palmer et al. (2000) who estimated the value to the user of seasonal forecasts for a range of cost:loss ratios. Although Palmer et al. (2000) considered a range of cost:loss ratios, the ratio was nevertheless considered to be fixed for any given user and mitigating action. The use of a fixed ratio may be appropriate to demonstrate the value of a forecast to, for example, a farmer who needs to decide whether to protect a crop from frost or perhaps even a particular country’s tourist industry that needs to decide whether to embark on a global advertising campaign. The DEMETER coupled model seasonal hindcasts give a maximum “value” (i.e., saving compared to the situation without a dynamical forecast) of \(\sim 15\%\) in winter or summer for the event “European average \(T_{2\text{m}} > \text{normal.}\) About half this value is achieved for the event that seasonal-mean precipitation is greater than normal. This value may be of interest to a utility company which is exposed to, and can take appropriate mitigating action against, European-wide climate anomalies.

Here, we explore the impact of predictability on a variety of possible mitigating actions. In particular, we consider the possibility that the cost:loss ratio may not be fixed for a particular mitigating action, but rather dependent on the forecast probability, \(p\), that the event will occur (\(p\) is not the climatological mean probability but the actual forecast probability for a particular date or time). Insurance is an example of an action where the cost:loss ratio may be a function of the forecast probability. This is because it is reasonable to assume that the insurance broker will also be aware of the probability and will charge a higher premium if the event is more likely to occur.

Suppose that a particular user is vulnerable to a particular weather or climate event. On average, this event occurs a fraction \(p_{\text{clim}}^t\) of the time. At any given time \(t\), a probabilistic forecast model predicts that the event will occur with probability \(p_t\). Let \(e_i = 1\) or 0, depending on whether the event happens or not in reality at time \(t\). We will assume that the forecast system is perfectly “reliable.” This means that \(E[e_i \cdot \alpha] = \alpha\) for all \(0 \leq \alpha \leq 1\) where \(E\) is the “expectation operator” (i.e., the long-term mean value). More intuitively, for a discrete set of possible forecast probabilities, this is equivalent to saying that “an \(\alpha \times 100\%\) probability forecast will verify \(\alpha \times 100\%\) of the time for any \(\alpha\) between 0 and 1.” Let \(q_i = 1\) or 0, depending on whether mitigating action was taken beforehand or not. Whether mitigating action is taken may depend on the forecast probability or just the climatological probability (see later). We will assume that any nonmeteorological influences on the user can be separated from the meteorological ones and are of no interest here (see later for a discussion). Let \(\Delta B_i\) be the change in the user’s cash balance associated only with the occurrence or not of the meteorological event at time \(t\). Here \(\Delta B_i\) will reflect the cost of taking action and the loss if no action is taken. Here we are interested in the behavior of the user’s accumulated cash balance, \(B_i(t_k) = \sum_{i=1}^{k} \Delta B_i\), for a range of possible mitigation strategies:

1) **Never take action.** Suppose that the user’s mitigation strategy is to never do anything. A loss, \(L\), is made if the event occurs \((e_i = 1)\) and no loss is made otherwise. Hence \(\Delta B_i = -e_i L\). The change in the user’s cash balance, the long-term expected change, and variance of change can be summarized as

\[
\Delta B_i = -e_i L \\
E[\Delta B_i] = -p_{\text{clim}} L \\
\text{Var}[\Delta B_i] = p_{\text{clim}}(1 - p_{\text{clim}}) L^2,
\]

where \(E[X]\) is the expected value of \(X\) (i.e., the long-term mean of \(X\)) and \(\text{Var}[X] = E[(X - E[X])^2]\) is the usual population variance. We will call \(-E[\Delta B_i]\) the “expected expense” associated with the user’s exposure to the given weather event.

2) **Always take action at fixed cost.** Suppose that the user’s mitigation strategy is to always take action at fixed cost \(C\) regardless of the forecast probability. The same quantities for this strategy are

\[
\Delta B_i = C \\
E[\Delta B_i] = -C \\
\text{Var}[\Delta B_i] = 0
\]

Traditionally, with no dynamic forecast information (only the climatological probability) it is assumed that the user will opt for whichever of these two above strategies has the smallest expected expense.
Hence if \( p_{\text{clim}} L < C \), then the strategy of never taking action would be preferred. However, the user’s accumulated cash balance for any of the strategies mentioned here is mathematically equivalent to a “random walk.” (Assuming that the weather or climate at time \( t_{i+1} \) is independent of that at time \( t_i \)). Hence the variance of the accumulated balance is given by

\[
\text{Var}[B(t_k)] = \sum_{i=1}^{k} \text{Var}[\Delta B_i] = k \text{Var}[\Delta B_1].
\] (14)

This means that the variance of the user’s cash balance increases linearly with time (if \( \text{Var}[\Delta B_1] \neq 0 \)). Since “volatility” is often an issue for the user (particularly if they are worried about bankruptcy, e.g.), it is possible that they may adopt the “always take action” strategy even if the expected expense is greater. The word “volatility” is used here as a generic term for variations about the expected value. For some users, volatility may be more important than expected expense and should be included in an assessment of the value of forecast information.

3) \textit{If sufficiently likely, take action at fixed cost.} If the user has access to a dynamical probabilistic forecast then they could decide to take action (at fixed cost) only if the forecast probability exceeds a given critical threshold, \( p_{\text{crit}} \). The same quantities for this strategy can be written as

\[
\Delta B_i = -(p_i > p_{\text{crit}}) C - (p_i \leq p_{\text{crit}}) e_i L,
\]

\[
E[\Delta B_i] = -C \left( 1 - \frac{C}{2L} \right),
\]

\[
\text{Var}[\Delta B_i] = \frac{C^2}{2} \left( 1 - \frac{C^2}{2L^2} \right),
\]

where \((x > y) = 1\) if the statement is true and 0 if it is false. The first term in \( \Delta B_i \) reflects the cost of taking action if the threshold is exceeded and the second term reflects the loss incurred if the threshold is not exceeded but the event nevertheless occurs. The expectation and variance calculations above assume a uniform probability density function (pdf) for \( p \). (For a discrete set of possible forecast probabilities, this would be equivalent to ensuring that each possible probability value was issued with the same frequency). Since we have assumed perfect reliability, a uniform distribution implies that \( p_{\text{crit}} = 0.5 \). As we have seen, such a value of \( p_{\text{crit}} \) is appropriate for the event “European \( T_{850} > \text{Normal} \).” Here \( p_{\text{crit}} \) has also been optimized to reduce the expected expense, giving \( p_{\text{crit}} = (C/L) \). It can be seen that the expected expense is less for this strategy than for the “always take action” strategy. This is because there is a saving made by not taking action when the forecast probability is small. However, there is a trade-off in terms of variance (which is now nonzero). It can be shown that for a uniform pdf (and \( p_{\text{crit}} = 0.5 \)), the strategy “if sufficiently likely, take action at fixed cost” is superior in terms of expected expense \textit{and} variance to the “never take action” strategy for all \( C < L \).

4) \textit{Always insure with knowledge of climatology only.} To reduce volatility, one often takes out insurance. If the user (and the insurance broker) has no access to a dynamical probability forecast, we can assume that the insurance premium will reflect the climatological probability of the event occurring. For this strategy we have (assuming the broker’s fees are negligible):

\[
\Delta B_i = -p_{\text{clim}} L,
\]

\[
E[\Delta B_i] = -p_{\text{clim}} L,
\]

\[
\text{Var}[\Delta B_i] = 0.
\]

Hence insuring with knowledge of only the climatological probability is similar to the “always take action at fixed cost” strategy (both lead to zero variance in the accumulated cash balance). The choice between these two strategies would simply depend on whether \( p_{\text{clim}} L < C \).

5) \textit{Always insure with knowledge of dynamical probabilistic forecast.} If a dynamical probability forecast is available, then the insurance premium may well reflect this. As stated earlier, it is reasonable to assume that the insurance broker also has access to the forecast. This strategy leads to

\[
\Delta B_i = -p_{i} L,
\]

\[
E[\Delta B_i] = -p_{\text{clim}} L,
\]

\[
\text{Var}[\Delta B_i] = \frac{L^2}{12},
\]

where the variance calculation has assumed a uniform pdf for \( p \) (and thus \( p_{\text{clim}} = 0.5 \)). The expected expense of insuring with a knowledge of the forecast probability is the same as when insuring with only knowledge of the climatological probability. However, the availability of a dynamical probabilistic forecast leads to volatility: the very aspect that insurance is designed to reduce. As before, the variance of the user’s cash balance will grow linearly with time. Hence forecast information diminishes the effectiveness of an insurance-based mitigation strategy in reducing a user’s cash volatility. One can conclude that the user who only has the option of taking out insurance will not benefit from dynamical forecast information.
Insurers themselves are often cited as major potential customers for forecast products, partly because of their familiarity with working with probabilistic information. However, if the forecast is universally available, one could argue that predictability does not improve the profitability of the insurance industry as a whole either. For example, in the limit of perfect deterministic predictability, either the user would not require insurance (if the event was definitely not going to happen) or the insurance premium would be at least as large as the unmitigated loss (if the event was definitely going to happen). Either way, insurance would not be sold. However, the insurance industry may not lose too heavily from predictability either as the main effect may be that the user begins to insure further ahead (beyond the limit of predictability). Although the use of forecast information in insurance may be inevitable in the long term, we could make the humorous comment that the only long-term “winners” associated with insurance-based actions may be the providers themselves! However, as a reviewer pointed out, insurance companies also want to reduce their own volatility and can do this by balancing their portfolio with anticorrelated “trades.” The use of forecast information may help identify such anticorrelated trades.

It is important to point out that, under the above assumptions of a uniform pdf (and \( p_{\text{clim}} = 0.5 \)), the variance of the user’s cash balance for the strategy of insuring with a knowledge of the dynamical forecast is exactly one-third that for the strategy of never taking action. Hence, although forecast information diminishes the effectiveness of an insurance-based mitigation strategy in reducing a user’s cash volatility, insurance may remain a better option than doing nothing.

Figure 8 illustrates the results of this section for the choice of cost \( C = 1 \), loss \( L = 2 \), and a uniform pdf. To model this scenario, for each time \( t \), we first choose \( p_t \) from the uniform distribution \([0, 1]\) and then choose a second random variable, \( q_t \), from the same distribution. If \( q_t < p_t \), then the event is deemed to happen \( (e_t = 1) \), if \( q_t \geq p_t \), then the event does not happen \( (e_t = 0) \). This process satisfies the reliability assumption and gives \( p_{\text{clim}} = 0.5 \). Using this process, two 200-element time series \([p_t]\) and \([e_t]\) are produced. The curves in Fig. 8 show \( B(t_k) + kC \) for each mitigation strategy for the same realization of \([p_t]\) and \([e_t]\). Adding \( kC \) to the accumulated balance ensures the curve for “always take action at fixed cost” (thick dotted line) is horizontal and allows easier comparison with the other mitigation strategies. The outcome of the “never take action” strategy is shown by the thick solid curve. For this strategy we also have that \( E[B(t_k) + kC] = 0 \) but the variance is nonzero. The theoretical \pm 1 \text{ standard deviation}

![Diagram](image-url)

**Fig. 8.** The effect over 200 time intervals on a user’s cash balance of different mitigation strategies against a particular climate “event.” See main text for details of the model used. (a) No action is ever taken (thick solid). (b) Action is always taken at fixed cost (thick dotted). (c) Action is taken at fixed cost if the dynamical forecast probability of the event exceeds the critical threshold (thick dashed). (d) Insurance is taken with only a knowledge of the climatological probability (also represented by the thick dotted line). (e) Insurance is taken out with a knowledge of the dynamical forecast probability (thick dot-dashed line). The thin solid lines show the theoretical standard deviation (\pm 1 \text{ SD}) of the balance for the case of never take action.
A reviewer pointed out that some weather or climate anomalies can have a good impact (say $G$) on a particular user rather than a bad impact (with loss $L$). However, this possibility is effectively covered by the same equations as those above if one redefines the event to be when the good anomaly does not occur. The fixed cost of taking action becomes $G - C$ and the loss is simply $G$ ($C$ is the actual cost involved in preparing to take advantage of the good anomaly). Note that if $G = 2$ and $C = 1$, then the curves in Fig. 8 would be exactly the same for this good weather anomaly.

The above analysis considers strategies that either involve a fixed cost or a cost that is proportional to the forecast probability that the event will occur. Strategies that lie somewhere between these two extremes may be more representative of the real world. For example, without a dynamical forecast, an intense cold spell over Europe might lead to a spike in gas demand and gas prices. A dynamical probabilistic forecast of this event may prompt a utility company to buy and store half the gas required for such an event beforehand and to insure against the other half. In an idealized situation, such a strategy can be represented as $\frac{1}{2} \times “$if sufficiently likely, take action at fixed cost” + $\frac{1}{2} \times “always insure with a knowledge of the dynamical probabilistic forecast.”” (Since any stored gas can be subsequently used whether or not the event happens, the fixed cost is associated with the cost of storage.)

It is interesting to note that, for the parameters used in Fig. 8, this strategy leads to less variance in the user’s cash balance than either of the two strategies alone. Wherever storage (of, e.g., water, gas, sugar, etc., but possibly not electricity) is an option, forecast information may well be able to reduce volatility. Another possible strategy that a user could consider may be to insure only if the event is sufficiently likely. However, this strategy actually leads to the same expected expense as the strategy “always insure with a knowledge of the dynamical probabilistic forecast.” With the uniform pdf assumption for $p$, the variance is minimized for a critical probability threshold of $p^{\text{crit}} = 0$. Hence if this strategy is optimized it is identical to the strategy “always insure with a knowledge of the dynamical probabilistic forecast.”

Energy traders apparently make the best profits from forecast information in the first few minutes of a new forecast being made available (I. Gilmour 2004, personal communication). This is presumably because energy prices initially reflect the predicted demand associated with the previous forecast. Within a few minutes, trading acts to adjust prices so that they reflect the new forecast. It is clear that some users will pay a premium to have first access to forecast information. Should such a market, which will lead to losers as well as winners, guide end-to-end forecast system development? It is also clear that consistency between forecasts made one day apart is an important issue for energy trading. Large discrepancies between two such forecasts will lead to volatility in energy prices.

The above analysis on mitigating actions is admittedly simple. Rational actions associated with “game theory” and irrational human behavior and other societal pressures will add considerable complexity to the real world. Phillips et al. (2002) looked at the potential impact of seasonal forecast information for crop production in Zimbabwe. In that study, the true complexity of the real world is very evident. Broad and Agrawala (2000) give a sobering account of the use of forecast information when set aside other societal pressures such as civil strife. While bearing these real-world complexities in mind, the present analysis does highlight several important issues that need to be considered when developing “end-to-end” forecasting systems: 1) Weather (or climate) forecasts can also affect the cost of mitigating actions. Depending on the nature of the mitigating action, this effect may need to be considered when estimating the “value” of a forecast system. 2) The value to a particular user of a forecasting system may depend on its ability to reduce volatility as well as to reduce expected expense. 3) The consistency of forecasts (how probabilities differ from one ensemble forecast to the subsequent ensemble forecast) is an important quantity to assess as it can have an impact on volatility. 4) The interests of different users are not necessarily aligned. Forecasters need to consider carefully which user communities to target when developing end-to-end forecasting systems. For example, should they target commodity trading, the agricultural policies of whole countries, or individual farmers?

5. Conclusions and key issues for the future

The extreme European summer heat wave of 2003 was used to introduce concepts of predictability and it showed anecdotal evidence of skillful prediction of European temperatures throughout the first month of a forecast. Medium-range forecast skill is well known to have increased substantially over the last few decades. For example, a “deterministic” 7-day ECMWF forecast of European 500-hPa geopotential height is as good today as a 5-day forecast was in 1980 (Simmons and Hollingsworth 2002; Lalaurette et al. 2003). What may be a more surprising result from our systematic “unified” analysis of operational ensemble forecasts is that there is indeed good probabilistic skill for European temperatures throughout the first month. For example, a Brier score of 0.16 is obtained for the prediction that...
monthly mean, European average temperatures at 850 hPa will be greater than normal. It was argued that this forecast, with short lead time, makes good use of the initial atmospheric conditions and, with a long averaging period, also benefits from a reduced unpredictable atmospheric component relative to the more sustained boundary-forced predictable signal. We have noted that the spread of an ensemble of coupled model forecasts can be increased, even in month 2, if atmospheric (not just sea surface temperature) perturbations are made to the initial conditions. In particular, this result suggests that the initial atmospheric conditions may be important for longer than has perhaps been previously recognized. It is unclear at present what role the coupled ocean plays in this spread increase. Our unified analysis results suggest that there is more probabilistic skill to be gained in the prediction of mean European temperatures averaged over (say) weeks 2–6 of a “seasonal” forecast.

Observationally based work does suggest a degree of seasonal predictability from the simple persistence of atmospheric anomalies. Anomaly correlations based on a multimodel ensemble mean suggest that much of this persistence is already captured in our models. If this is the case, then improvements in skill may rely on our ability to predict the evolution, rather than simple persistence, of anomalies. (We have noted however that the seasonal prediction of the 2003 summer heat wave was sensitive to the amount of soil moisture in the initial conditions, so there may be more skill to be gained from decayed persistence for such extreme events.) The role of oceanic SST in forcing European seasonal climate variations remains unclear. While analysis of variance suggests a rather weak general effect for Europe, there is some observational evidence that particular SST anomaly patterns may have a relatively strong impact on European temperatures. Idealized atmospheric model studies forced with these particular SST anomaly patterns tend to confirm this impact and, if anything, suggest that present models underestimates the true signal. Hence there could be “windows of opportunity” when European predictability is enhanced. Model validation and improvement may help forecasts tap better this source of predictability. Similar “windows of opportunity” could equally arise associated with strong land surface or possibly stratospheric anomalies and may also be a function of the time of year. Again, more model validation work is required.

The use of multiple-model seasonal ensemble forecasting systems is gaining popularity at present. These multimodel systems offer improved probabilistic skill scores, partly because of an improved sampling of uncertainty. A three-model system, comprising the same three coupled models discussed here, is shortly to become operational at ECMWF. In the longer term, it is possible that a “single”-model system could again be favored but with ensemble-member perturbations including perturbations to the physical parameterizations that reflect present uncertainty in the science (Murphy et al. 2004; Palmer et al. 2005). (This is different and complementary to the “stochastic forcing” associated with uncertainties in unresolved processes that is presently applied to the tendencies in the ECMWF ensemble system.)

We have attempted to estimate ultimate levels of predictability (and thus goals to which we can aim for) by calculating coupled model “potential predictability.” This is done by assuming that a particular coupled model is perfect, using one ensemble member to represent the truth and investigating how well the other ensemble members are able to predict this “truth.” However, potential predictability estimates appear to be model dependent. It may be that present models are too imperfect for us to be able to assume they are perfect! Further effort is required to quantify coupled model potential predictability and its sensitivity to the model used.

The utility or otherwise of monthly to decadal forecasts for Europe may rely on careful optimization of the whole “end-to-end” forecast-to-user decision-making process. Important issues include the recognition, in estimates of forecast “value,” that forecasts can affect the cost of mitigating actions, and that reducing volatility (not just expense) may be an important motivator for some users. Since the interests of different users are not necessarily aligned, careful consideration may be required of which user communities to target when developing end-to-end forecasting systems.

Acknowledgments. The authors thank Renate Hagedorn, Michel Déqué, Hervé Douville, Laura Ferranti, Claude Frankignoul, Thomas Jung, Robert Mureau, Tim Stockdale, Rowan Sutton, and Geert van Oldenborgh for useful discussions about this paper. The authors would also like to thank the reviewers for their useful and insightful comments. FJDR was supported by the EU-funded ENSEMBLES project (GOCE-CT-2003-505539).

REFERENCES


CORRIGENDUM

It has come to our attention that there is a typographical error in “Medium-Range, Monthly, and Seasonal Prediction for Europe and the Use of Forecast Information,” by Mark J. Rodwell and Francisco J. Doblas-Reyes, which was published in the Journal of Climate, Vol. 19, No. 23, 6025–6046. In section 4 on p. 6040, a spurious term \( \sqrt{b^2 - 4ac} \) was not deleted from the text in the first mitigation strategy listed. The correct text should have appeared as follows:

1) Never take action. Suppose that the user’s mitigation strategy is to never do anything. A loss, \( L \), is made if the event occurs \( (e_i = 1) \) and no loss is made otherwise. Hence, \( \Delta B_i = -e_iL \). The change in the user’s cash balance, the long-term expected change, and variance of change can be summarized as

\[
\Delta B_i = -e_iL,
\]

\[
E[\Delta B_i] = -p_{\text{clim}}L,
\]

\[
\text{Var}[\Delta B_i] = p_{\text{clim}}(1 - p_{\text{clim}})L^2,
\]

where \( E[X] \) is the expected value of \( X \) (i.e., the long-term mean of \( X \)) and \( \text{Var}[X] = E[(X - E[X])^2] \) is the usual population variance. We will call \(-E[\Delta B_i] \) the “expected expense” associated with the user’s exposure to the given weather event.

The staff of the Journal of Climate regrets any inconvenience this error may have caused.