Predictability of Northeast Brazil Rainfall and Real-Time Forecast Skill, 1987–98

CHRIS K. FOLLAND, ANDREW W. COLMAN, DAVID P. ROWELL, AND MIKE K. DAVEY
Hadley Centre for Climate Prediction and Research, Met Office, Bracknell, Berkshire, United Kingdom

(Manuscript received 7 February 2000, in final form 20 July 2000)

ABSTRACT

The predictability of rainy season rainfall over northeast Brazil for the relatively long period 1912–98 is analyzed using dynamical and empirical techniques. The dynamical assessments are based on the HadAM2b atmospheric model forced with the Met Office Global Sea Ice and Sea Surface Temperature Dataset (GISST3). Ensembles of simulations and hindcasts starting from real initial conditions for 1982–93 made under the European Community Prediction of Climate Variations on Seasonal to Interannual Timescales (PROVOST) program are analyzed. The results demonstrate a relatively high degree of predictability. Its source lies mostly in tropical Atlantic and Pacific sea surface temperatures. The results confirm the less extensive evidence of other authors that northeast Brazil is a region where two separate ocean basins influence seasonal climate to a comparable extent. Overall, the sea surface temperature gradient between the northern and southern tropical Atlantic appears to be the more important influence, though El Niño can be dominant when it is strong. These assessments of predictability are consistent with the performance of over a decade of real-time long lead and updated forecasts, issued over the period 1987–98. Multiple regression and linear discriminant analysis prediction techniques, together with model forecasts in the last few years, were used to provide best estimate and probability real-time forecasts of rainy season rainfall. These forecasts had a level of skill that was close to the state of the art in seasonal forecasting.

1. Introduction

The north-northeast region of Brazil (Fig. 1), henceforth called northeast Brazil (NEB), receives about 60% of its annual rainfall during March–May when the intertropical convergence zone (ITCZ) moves closest to the region. The dynamical background is discussed in Hastenrath (1995a). Much work has been carried out to understand or predict NEB rainfall. Walker (e.g., 1928) may have been the first to suggest a possible link between the Southern Oscillation and NEB rainfall and to estimate its predictability (Walker 1930). Markham and McLain (1977) and Hastenrath and Heller (1977) presented observational evidence of a possible relationship between Atlantic sea surface temperature (SST) anomalies and NEB rainfall extended to the east Pacific by Hastenrath (1978). Moura and Shukla (1981) presented evidence of a mechanism linking a dipole of tropical Atlantic SST north and south of the equator and NEB drought from global and local atmospheric general circulation model (AGCM) experiments. Mechoso et al (1990) also found evidence of an influence of tropical Atlantic SST anomalies on NEB rainfall.

Several AGCMs have been shown to reproduce well NEB rainfall (e.g., Livezey et al. 1995; Potts et al. 1996, hereafter PFJS; Harzallah et al. 1996; Sperber and Palmer 1996). Ward et al. (1988), Ward and Folland (1991, hereafter WF), and Ward et al. (1993) have documented strong, apparently linear, relationships between Pacific and tropical Atlantic SST anomaly patterns and seasonal rainfall in NEB, results updated and summarized in Carson (1998) and Ropelewski and Folland (1999). This relationship is the basis of 12 largely successful sets of real-time forecasts of NEB issued by the Met Office annually since 1987. Hastenrath has also produced real-time forecasts for recent rainfall seasons (Hastenrath and Greischar 1993b). These forecasts are the result of several years of research (e.g., Hastenrath et al. 1984; Hastenrath 1992). Hastenrath’s group uses preseason rainfall and meridional wind components as well as SST as predictors, whereas we have used SST alone.

We first discuss the rainfall data in section 2. We then review the methods we use to assess predictability and real-time forecast skill in section 3. Section 4 discusses the observed relationships between NEB rainfall and SST; these are exploited in section 5, which discusses empirical methods for assessing predictability. Section 6 discusses the dynamical predictability of NEB rainfall using an atmospheric general circulation model forced with observed SST while section 7 discusses the performance of over a decade of real-time forecasts.
paper ends with a discussion of these results in the light of the estimates of potential predictability followed by conclusions in section 9.

2. Rainfall data

Three indices of seasonal NEB rainfall have been used for real-time forecasts. They are 1) an index of March–May rainfall for two key stations, Fortaleza and Quixeramobim (FQ) shown in the Fig. 1 insert; 2) an index of March–April rainfall for 27 stations for 1912–81 marked by `*` in Fig. 1, compiled by Hastenrath; and 3) an index of February–May rainfall for 1912–81 for over 100 stations supplied by C. Nobre (1988, personal communication), representing the area bounded by the thick solid line in Fig. 1. An almost complete dataset of monthly rainfall values for 1912–94 for the 27 stations marked by * in Fig. 1 was also supplied by Hastenrath. This dataset was updated to 1998 using data supplied by the Brazilian Centro de Pesquisa de Tempo e Estudos Climáticos (CPTEC) and by Fundação Cearáense de Meteorologia e recursos hidricos (FUNCEME) in Fortaleza, Brazil. These monthly data were used to update the Hastenrath March–April seasonal index. Gridded data supplied by Hulme (1994) for the six 2.5° lat × 3.75° long boxes outlined in bold in Fig. 1 were utilized to update the Nobre index and to verify rainfall simulations outside the NEB area.

For investigating the predictability of NEB rainfall, further indices were created from the Hastenrath monthly data. To avoid biases due to a time-varying station density because of data gaps, station values in the Hastenrath set were interpolated on to grid points at 1° lat and 1° long intervals (crosses in Fig. 1 inset). Standardized gridpoint values were calculated by averaging weighted standardized values for all the stations within an ellipse whose axes were 1° lat and 1° long from the grid point. The weights were calculated as the reciprocal of the distance between the grid point and the station. A simple average of the 27 gridpoint series provides the March–May NEB rainfall index. Other grids gave very similar rainfall indices.

Climatological averages are calculated over 1951–80 for continuity with past work. The timing of the rainfall season in NEB is not sharply defined. For the Hastenrath stations the wettest months are March (194 mm) and April (161 mm), together accounting for just under half the annual total (746 mm). Fortaleza is the wettest of the 27 Hastenrath stations; thus the FQ totals are higher (240 mm in March, 253 mm in April) with the rainfall season extending into May (182 mm). However, the interannual variability is largely independent of the choice of months. For example, the correlation between the Hastenrath series for February–May and March–April over the period 1912–95 was 0.90, while that between February–May and March–May averages was 0.94. In this paper we investigate predictability for periods of varying length, with some emphasis on March–May. The spatial coherence of the rainy season rainfall totals over the region is very strong as shown by an eigenvector analysis of March–May rainfall for the Hastenrath region for 1912–95. The first eigenvector has almost equal weights (not shown) and explains 70% of the seasonal rainfall variance. In addition there is a very high correlation \((r = 0.98)\) between the February–May Nobre series and a February–May series produced from the Hastenrath stations.

3. Methods of assessment

In the following sections, the performance of SST-based predictors for forecasting or simulating NEB rainfall is discussed. First, we discuss predictability from SST using empirical methods and AGCM runs. To measure predictability, the following measures are used:

1) Pearson, or standard, correlation in timewise mode.
2) Squared coherence (Bloomfield 1976), which measures the similarity of time series over different frequency ranges.
3) Linear Error in Probability Space (LEPS) in timewise mode (PFJS).

Standard correlation and squared coherence do not take account of bias between two series or difference in standard deviation, but knowledge of these is not always essential for assessing predictability though they must be compensated for in real-time forecasts. However, significant biases and differences in standard deviation are pointed out when they occur and LEPS is used sometimes to highlight this.

LEPS measures an error of a forecast as the “distance” in a chosen climatological cumulative probability distribution (most often of the verifying observations) between a forecast and the corresponding observation, referred to the chance distance created by random forecasts (PFJS). Thus if the error in the forecast...
was exactly equal to the chance error, the LEPS score would be zero. LEPS comes in several forms. For measuring the skill of point estimates we use the continuous version, LEPSCONT, and for categories we use LEPSCAT, which is merely a discrete form of LEPSCONT. We use quint (five equiprobable) categories based on Table 2 of PFJS, but with LEPS scores converted to percentage skill scores in the range +100% to −100% as in section 7 of that paper. Two versions of LEPSCONT or LEPSCAT are used (a) LEPSOB, which uses the probability distribution of the observations to rank the observations and forecasts (b) LEPSOBFC where we use the probability distribution of the observations to rank the observations and that of the forecasts to rank the forecasts. LEPSOB is sensitive to forecast bias and errors in forecast standard deviation, whereas LEPSOBFC is not, and behaves more like standard correlation. Where forecast or simulation data are standardized, LEPSOB and LEPSOBFC give practically the same values and LEPSOB is used.

Second, for measuring the skill of real-time best estimate forecasts we also use (Nicholls 1984):

1) root-mean-square error (rmse).
2) Bias (difference between mean of forecasts and mean of observations).

For measuring the skill of real-time probability forecasts made by linear discriminant analysis, the following measures are used:

1) LEPSPROB skill. This is like LEPSCAT in that it uses the same table of weights to calculate a score and a skill from the forecast/observed contingency table. However, instead of a single forecast category being given a weight of one, each forecast category is given a fractional weight according to its forecast probability (WF).
2) RPS skill (Ranked Probability Score; Epstein 1969).

Our standard period for estimating predictability is 1948–97 as SST data is best at that time, though most analyses were extended back to 1912 with similar results unless stated otherwise. The main exception is the assessment of the Prediction of Climate Variations on Seasonal to Interannual Timescales (PROVOST) hindcasts and simulations, which were made for rainfall seasons between 1982 and 1993. In addition, real-time forecast skill for 1987–98 is selectively compared with hindcast skill over the period 1912–86.

4. Correlations of northeast Brazil rainfall with SST

The Global Sea Ice and Sea Surface Temperature version 3 (GISST3) dataset (updated from Rayner et al. 1996) allows relationships between NEB rainfall and SST described by earlier authors to be reassessed. GISST3 is a globally complete dataset, consisting of smoothed observed data blended with data reconstructed using eigenvectors and Laplacian functions. The eigenvector reconstructions are very effective at reconstructing El Niño SST variations in data-sparse periods like the early twentieth century. Most previous studies relating NEB rainfall and SST have been made using less complete SST data, for example, Markham and McLain (1977), Hastenrath and Heller (1977), Hastenrath (1978), Hastenrath et al. (1984), WF, Sperber and Hameed (1993), and Uvo et al. (1998).

Simultaneous correlations between Hastenrath rainfall and worldwide SST for 1948–97 are shown in Figs. 2a–d for the four two-month periods January–February, March–April, May–June, and June–July. Other periods are not shown because the rainfall is generally low. SST values are calculated over near-equal areas of average size 10° lat × 12° long at the equator and shaded areas are locally significant at the 95% confidence level. Such areas reflect the greater spatial correlation of SST in an east–west than a north–south direction. Areas where GISST3 uses Laplacian interpolation instead of eigenvector reconstructions (due to lack of data) are omitted.
as reconstructed interannual variations of SST in these regions are likely to be too weak. In January–February (Fig. 2a), the fraction $f$ of global area analyzed that is locally 95% significant is $f = 0.10$. In Fig. 2a, the dipole of tropical Atlantic correlations gains some strength compared to a weak pattern in December–January (not shown), though most of the strongest (positive) correlations are in the south Atlantic. South of 20°S, correlations become weakly negative. The well-known negative correlation of NEB rainfall with the El Niño pattern in the Pacific is clear but weak. Relationships among NEB rainfall, El Niño–Southern Oscillation (ENSO), and the north–south Atlantic SST dipole are well documented (Hastenrath 1984, 1993; Ward and Folland 1991; Moura and Shukla 1981 and many others). In March–April ($f = 0.38$) the correlations have the same pattern but are much stronger, though the largest local values are in the tropical South Atlantic while the negative correlations to their south start to retreat. The negative correlations in the Indian Ocean reflect the influence of El Niño on SST there. Correlations are a little weaker in the tropical Pacific in April–May (Fig. 2c) than in Fig. 2b, but are just as strong in the Atlantic, and stronger near the South American coast ($f = 0.34$). Correlations generally weaken in May–June (not shown) but it is interesting to note that in June–July (Fig. 2d, $f = 0.12$) when the influence of ENSO is low, a residual but distinctive pattern of positive correlations persists in the South Atlantic off the coast of Brazil down to at least 30°S. This may hint at influences of the South Atlantic Convergence Zone. Thus if that zone is moved north by warm SSTs around 10°–20°S where peak correlations occur, the end of the rainfall season may be delayed (I. Cavalcanti 1997, personal communication). Such effects are worth investigating.

In the dry season, correlations are very weak. They are essentially zero in the tropical east Pacific between July–August and September–October, or slightly positive in August–September. In these periods the Atlantic SST dipole has no obvious skill. By October–November, the main rainy season SST correlation pattern is re-established weakly.

Although correlations of SST with rainfall in individual months underlying Fig. 2 are subject to more noise and are not illustrated here, they are worth remarking on. A curious feature is a weakening of simultaneous correlations as the early NEB rainfall season progresses. The correlation between January SST and January NEB rainfall is stronger than that between February SST and February NEB rainfall in both the tropical Pacific and tropical Atlantic, the opposite of that expected, as noticed by Uvo et al. (1998). Correlations strengthen sharply in March over the tropical Pacific and north tropical Atlantic. Further strengthening occurs in all regions in April with some weakening in May. June shows much less simultaneous correlation, except in the south tropical Atlantic though, unexpectedly, some increase in correlation occurs in the tropical Pacific in July. Although they carried out their rainfall analysis over a larger area, the strong peak of correlation of SST in April, reducing somewhat in May, is consistent with the singular value decomposition results of Uvo et al. (1998). However, our results generally show a larger extent of ocean with locally significant correlations. This may reflect the better signal to noise ratio characteristics of our larger equal area SST boxes.

Figures 3a–d show correlations of NEB rainfall with SST in the month but one preceding the start of each period for the four two-month rainfall periods January–February, March–April, April–May, and June–July. Thus for rainfall in January–February, the relevant SST month is the previous November. These maps indicate the relationships that are most useful for long lead seasonal rainfall forecasts. For January–February ($f = 0.12$) there is little evidence of a tropical Pacific correlation (Fig. 3a). In the Atlantic, only negative correlations with SST in the tropical north Atlantic are significant. By March–April (Fig. 3b, $f = 0.33$), a strong negative tropical Pacific correlation is evident with sig-
significant positive correlations in the South Atlantic and weaker negative correlations in the tropical north Atlantic. This is much like the simultaneous correlations in Fig. 2b. In April–May (Fig. 3c, $f = 0.35$), the tropical north and south Atlantic correlations strengthen while the tropical Pacific correlations weaken, but remain widely significant. By June–July ($f = 0.16$) correlations in the tropical Pacific are weak, though still negative, but those in the tropical Atlantic remain strong, indicating some predictability for the late rainfall season.

Correlation patterns are much the same when the analysis period is shortened to 1958–97. Consequently, correlations for 1912–57 and 1958–97 have similar spatial patterns (not shown), indicating that the seasonally varying relationships between NEB rainfall and SST were nearly independent of global warming during the twentieth century.

5. Use of sea surface temperature eigenvectors for empirical prediction

Ward and Folland (1991) used two periods (1901–80 and 1949–98) to calculate covariance eigenvectors (EOFs) of SST at $10^\circ$ lat $\times 10^\circ$ long resolution for forecasting NEB rainfall. They considered the 1901–80 EOFs to be most useful for prediction. They used just two of these EOFs. They are (i) EOF3 of all seasons Atlantic SST (Fig. 4a), which has centers of high weight roughly corresponding to the centers of significant correlation in Figs. 2 and 3 in the tropical Atlantic and (ii) EOF1 of December–February Pacific SST (Fig. 4b), which is strongly ENSO related. Atlantic EOF3 also includes a dipole pattern of correlation of SST with NEB rainfall within the south Atlantic. Thus SST is generally negatively or weakly correlated with NEB rainfall south of $20^\circ$S in the South Atlantic (as in Figs. 2 and especially 3). Ward and Folland (1991) only use the tropical ($30^\circ$N–$30^\circ$S) part of Atlantic EOF3, which still includes some of this south Atlantic SST dipole but use all of the tropical Pacific EOF. We refer to these as the WF EOFs. A pre-GISST dataset, the Meteorological Office Sea Surface Temperature Data Set version 3 (MOHSST3), which had less data coverage and a poorer data gap filling mechanism was used to compute the WF EOFs. The time series of the EOFs (scalar products between SST anomaly fields and the EOF field) measure the strength of a particular EOF at a given time and can be used as a predictor index in statistical prediction. The two WF EOFs are not in principle orthogonal but in practice their March–May time series for 1912–97 had a correlation of only 0.01. These EOFs have been used for all real-time forecasts issued since 1990. Attempts to improve on these patterns as predictors by EOF analysis of various versions of GISST or canonical correlation analysis between the Hastenrath stations and GISST EOFs have produced comparably skillful prediction methods, but none with better skill.

Predictability from WF EOFs

Here we investigate the skill of the WF EOFs in simulating or hindcasting the Hastenrath rainfall series throughout the year. We use the term “simulation” when using predictor SSTs measured simultaneously with the rainfall and “hindcast” when the predictor SST is measured before the rainfall season begins, and real-time forecasts are not being considered. We assess the simulation skill of the WF SST EOFs for periods of the year for which the correlation analyses suggest that the EOFs are useful. For most of these assessments, standard correlation and LEPSOB or LEPSOBFC are used. In most cases, the variation in LEPS scores is similar to the correlations, except that LEPS skill values are generally not far from 0.7 times the correlation value. To minimize artificial skill, the year being hindcast or simulated, and that preceding and following, are excluded from the regression calculation. This is called the “jackknife” method. The number of years before and after the target year that should be excluded depends on the serial correlation of the predicted series. The
Fig. 5. Assessments of linear regression simulations and hindcasts of Hastenrath rainfall for 1948–97. The time series of the Fig. 4 EOFs are used as predictors; skill is measured by the standard correlation between simulated or hindcast and observed rainfall for periods between 1 and 6 months in length. (a) Simulation skill, 1948–97. (b) Hindcast skill, 1948–97, using predictors based on SST in the month but one preceding the beginning of the hindcast period. (c) Hindcast and simulation skill for Mar–May rainfall 1948–97 using predictors based on SST measured over one month, starting the previous Jun onward. Hindcast skill using the Tropical Atlantic EOF and tropical Pacific EOFs separately is also shown. (d) Hindcast and simulation skill for Mar–May rainfall using both predictors, but using SST averaged over periods 1, 2, or 3 months long ending at the month shown.

above choice is sufficient as NEB rainfall shows little interannual persistence (first serial correlation 1912–97 = 0.13); a quasi-biennial peak in the NEB rainfall spectrum (Chu 1984) contributes to this behavior. Thus a hindcast for 1960 would utilize a regression equation calculated using data for 1912–58 and 1962–97. This method gives as many regression equations as years being hindcast or simulated; the coefficients of the predictor variables in the regression equations also vary with time of year.

GISST3 is used to calculate the time series of the WF EOFs, though the EOFs themselves were calculated from MOHSST3. Figure 5a shows correlations between simulated and observed Hastenrath rainfall for periods of 1–6 months in length over 1948–97. Correlations are highest during the rainy season and generally increase with the length of simulation period. For shorter periods, correlation is a maximum for periods centered around March and April, the height of the rainy season, but for longer periods this shifts slightly to periods centered on February. The skill for individual months is generally poor with the strong exception of April ($r = 0.66$). Skill improves considerably for two-month periods so that March–April and April–May have good skill ($r = 0.62, 0.63$). Three-month periods are better still while six-month skill has correlations >0.6 from November–April to February–July. The corresponding LEPSOB scores have a similar shape but are considerably lower. LEPSOB penalises the assessments when the forecasts have a different distribution from the observations whereas LEPSOBFC does not. LEPSOB is generally lower than LEPSOBFC because the simulations have a smaller standard deviation than the observations. However, for six-month periods there is only a small difference in the standard deviations so that LEPSOB = 0.44 and LEPSOBFC = 0.49.

Figure 5b assesses hindcasts made from predictors based on SST in the month but one before the start of the hindcast period. This is a practical lead period for long lead forecasts. Remarkably, such hindcasts are often more skillful than the simulations in Fig. 5a. This may reflect a delay in the influence of SST on NEB rainfall. If true, this has beneficial implications for seasonal predictability here and possibly elsewhere in the Tropics. Maximum correlation scores for 1- to 6-month periods are in the range 0.74–0.78. April is again the
most skillful single month with a correlation as high as \( r = 0.78 \). The most skillful three-month period is April–June \( (r = 0.75) \) (note this is not the wettest period), the most skillful 4-month period is April–July \( (r = 0.74) \), and the most skillful 6-month period is April–September \( (r = 0.75) \). This is a strong indication that useful forecasts are possible beyond May for periods greater than 3 months long, though the WF EOFs may not be optimal. The LEPSOB and LEPSOBFC skills are also higher than the corresponding LEPS skill of the simulations. Again LEPSOBFC has higher skill (typically by 0.1) than LEPSOB, with values in the range 0.50–0.55 for the most skillful periods 1–6 months long (not shown). From these results, it seems best to use inflated regression in real-time forecasts, which (nearly) achieves equality of forecast and observed variance. This has only been done for real-time forecasts from 1999.

Staying with the idea of using SST from a single month to create the prediction equation and concentrating on the core rainfall season March–May, Fig. 5c shows the correlation skill for hindcasts using the Atlantic and Pacific EOFs separately and combined. Thus, Fig. 5c also shows how skill varies as the delay between SST and the core forecast period increases. SST is measured in individual prior months from the June of the previous year to the last month of the forecast period, May. We first discuss the skill obtained from using Pacific and Atlantic EOFs together. Figure 5c provides additional guidance to that in WF (e.g., their Fig. 8) on how far ahead it is possible to forecast the main rainfall season. Taking a minimum correlation of 0.6 as clearly useful. March–May rainfall is skillfully predictable from SST measured in December onward, though November might be regarded as giving a useful preview. This corresponds to the real-time practice for preliminary forecasts discussed in this paper where SST averaged between November and January was often used. For January SST, the hindcast correlation is 0.65 (as in Fig. 5b, three-month curve), but peaks for February SST at \( r = 0.71 \). This result indicates that “updated” forecasts, or nowcasts, made in early March for March–May rainfall using February SST may give modest extra guidance, as has been done in real-time forecasting since 1988. Note how the skill declines when SST from within the forecast period is used.

Considering the Atlantic and Pacific Oceans separately, consistent with Eq. (4) of WF, the Atlantic WF EOF shows considerably more skill than does the Pacific EOF. Assuming 45 degrees of freedom, the Pacific EOF on its own does not have significant skill \( (r \approx 0.29) \) until February. Nevertheless, Fig. 5c shows that when combined in a stepwise linear regression equation, the combined skill of the two oceanic predictors, measured by linear correlation, exceeds that from either predictor individually after November. Thus NEB provides an example of predictability in a region where El Niño SST patterns play a significant but lesser role compared to a (nearly) independent SST pattern in another tropical ocean (the Atlantic). However, a small part of this Atlantic influence is an indirect impact of ENSO because ENSO affects tropical Atlantic SSTs (Curtis and Hastenrath 1995; Enfield and Mayer 1997; Sutton et al. 2000). Before November, on average almost all the skill for predicting the following March–May rainfall comes from the Atlantic SSTs, though this may not be true in some strong El Niño years.

Harzallah et al. (1995) suggested that NEB rainfall is usefully predictable from SST six months in advance. Our results indicate that this is perhaps not quite achievable for the core rainfall season March–May. However, an issue is the optimum period over which to average SST. So far we have not used averaging periods of less than one month (e.g., using the last week in January to make a long lead forecast in early February), believing that noise in the data would increase too much. [This may be less of a problem with reduced noise SST analyses now becoming available (Jones et al. 2001) than currently these are only made monthly]. So we have concentrated on finding out if averaging periods of two or three months have any benefit, as has often been used in the real-time forecasts. Any benefit is likely to come from a further reduction on noise, offset by the inclusion of older SST data.

Figure 5d shows how correlation skill varies with hindcast lead time and with the length of the period from which the SST predictors are calculated. The hindcast period is again fixed at March–May, and SST is measured over a variety of prior periods ending at the month shown. Only the results for the combined Atlantic and Pacific WF predictors are shown. Other methods of averaging the SST could be imagined and have been used in some real-time forecasts, for example, by giving more weight to the more recent months of a three-month averaging period. However, it is clear from Fig. 5d that there is no advantage in averaging SST beyond a month, though one and two months give similar skill, and three months marginally worse skill except for hindcasts using SST ending in October. Figure 5d indicates again that it seems inadvisable for March–May forecasts to be based on predictors using SST measured before November, though there is some skill. Additional information can be obtained by taking a closer look at the probability distribution of the hindcasts (Table 1). To make this useful for the future, we have used inflated regression as now used routinely in the real-time forecasts. Table 1 shows the skill of hindcasts in the five equiprobable (quint) categories that are routinely made using LEPSOB, which gives similar results to LEPSOBFC as inflated regression is used. A value greater than about 0.5 represents a very useful level of skill while 0.3 is marginal. Table 1 shows the well-known result (Van den Dool and Toth 1991) that the extremes have higher skill than the central categories, and that the average category has no obvious skill. Hindcasts using November SST have useful skill if a hindcast is...
made for an extreme quint, but not otherwise. Marginally useful skill perhaps exists for hindcast quints 2 and 4 in January and more definitely in February. Thus we must expect the bulk of the skill in seasonal forecasts for NEB to reside in the extremes but not to be solely confined to them. This is fortunate as the wet and dry extremes are societally the most important anomalies to predict.

Finally, we note that although Fig. 4a shows a dipole of tropical Atlantic SST anomalies as a major component of its pattern, the two halves of the dipole fluctuate nearly independently (e.g., Houghton and Tourre 1992; Enfield and Meyer 1997) except on near decadal (Mehta and Delworth 1995; Chang et al. 1997) and perhaps multidecadal (e.g., Folland et al. 1986) timescales. Nevertheless, Fig. 4a reflects the importance of the tropical interhemispheric SST gradient to NEB rainfall, which results from largely independent SST fluctuations in the two tropical hemispheres, of especially large amplitude (not shown) in the NEB rainy season.

6. Model simulation skill

Simulations of mean rainfall for NEB have been produced using the Met Office HadAM2b [Hadley Centre Atmospheric Model version 2b, Hall et al. (1995)] dynamical model forced with GISST3. The model resolution is 2.5° lat × 3.75° long with grid points in the NEB region at the center of the boxes in Fig. 1. These simulations update those discussed by PFJS and an ensemble size of six integrations was used. Compared to the PFJS model, the newer model has improved convection, surface hydrology, boundary layer, and gravity wave drag schemes. A much-improved SST anomaly reconstruction technique was used to fill gaps in GISST3 compared to GISST1.1 used by PFJS, so that some historical ENSO events are much better represented in GISST3 than in GISST1.1.

We have used model rainfall averaged over four model grid points to represent NEB rainfall. These grid points best represent the Hastenrath region (Fig. 1); they are 5°S, 37.5°W, and 41.25°W, and 7.5°S, 37.5°W and 41.25°W. Grid boxes at 2.5°S are situated over the ocean so are best not used (see remarks under seasonal prediction). The annual cycle of the model’s 1951–80 rainfall climatology has a very similar shape to that observed calculated from the Hastenrath station data but the model is somewhat wetter than observed in all months. The percentage difference ranges from 30 mm (15% wetter than observed) in March to 10 mm (100% wetter than observed) in the very dry month of October.

a. Model correlations with SST

Figure 6 shows simultaneous correlations between the average rainfall simulated by HadAM2b at the four grid

---

**Table 1. Skill of Mar–May hindcasts (1948–97) using SST predictors measured in selected preceding months and inflated regression.**

<table>
<thead>
<tr>
<th>SST period</th>
<th>Correlation</th>
<th>All</th>
<th>Driest 20%</th>
<th>20%–40%</th>
<th>40%–60%</th>
<th>60%–80%</th>
<th>Wettest 20%</th>
<th>Quint 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb</td>
<td>0.71</td>
<td>0.43</td>
<td>0.53</td>
<td>0.35</td>
<td>0.01</td>
<td>0.30</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Jan</td>
<td>0.65</td>
<td>0.39</td>
<td>0.63</td>
<td>0.20</td>
<td>0.08</td>
<td>0.37</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td>0.59</td>
<td>0.33</td>
<td>0.57</td>
<td>0.11</td>
<td>−0.07</td>
<td>0.23</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Nov</td>
<td>0.52</td>
<td>0.32</td>
<td>0.58</td>
<td>0.11</td>
<td>0.15</td>
<td>−0.01</td>
<td>0.67</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 6. Simultaneous correlation between ensemble mean HadAM2b rainfall and GISST3, 1948–97, in (a) Jan–Feb, (b) Mar–Apr, (c) May–Jun, and (d) Nov–Dec. The contour interval is 0.125, dashed contours are negative, and correlations for shaded areas are locally significant at the 95% confidence level.**
points and SST for 1948–97, calculated for the same months as the observations in Fig. 3. These correlations provide guidance on how the SST patterns influence the model in simulating NEB rainfall. In January–February (Fig. 6a) the pattern is similar to the observations in Fig. 3a except that correlations with the tropical central and eastern Pacific are larger and more significant. This may in part result from the use of ensemble mean data for the model rainfall time series, which removes a large part of the random effects of internal atmospheric variations. Both model and observations show only weak correlations with the north tropical Atlantic. In March–April (Fig. 6b), the model shows strong negative correlations with the tropical north Atlantic and the tropical Pacific and positive correlations with the tropical south Atlantic, similar to observations. The negative correlations with the tropical Pacific are again somewhat stronger than observed, though not so strikingly as in January–February. In May–June (Fig. 6c) correlations weaken in all three ocean basins, but only a little, and less than observed (not shown). At the beginning of the rainfall season (November–December, Fig. 6d), the model shows stronger negative correlations in the tropical central and east Pacific than observed, and stronger positive correlations in the south Tropical Atlantic, though like the observations, only very weak negative correlations are seen in the tropical North Atlantic. Overall, the strengthening of model correlations compared to the observed is mostly in the tropical Pacific. So, in linear terms at least, model rainfall is rather more influenced by tropical Pacific SSTs, and therefore ENSO, than are the observations but the effect is not large. An overstrong influence of ENSO on Atlantic sector tropical climate is a known problem with coupled models with similar atmospheric physics to HadAM2b (S. Tett 1998, personal communication), though NEB is not in the worst affected region.

b. Model simulations of Northeast Brazil rainfall

Figure 7a shows the skill of the model in simulating rainfall on timescales from 1–6 months (we emphasize these are not hindcasts). On the 1-month timescale, the model has positive skill in every month but only in April–July. On the 2-month timescale $r > 0.6$ from January–February until July–August reaching 0.80 in April–May. On the 3-month timescale $r$ is slightly higher again most of the year, exceeding 0.7 from January–March until June–August and peaking at $r = 0.84$ in April–June. Six-month periods show generally the highest skills, all but two 6-month periods (out of 12 possible) having $r > 0.6$. Peak skill occurs around periods centered on April and May when $r = 0.87$. Thus the model has exceptional potential when skill is measured this way and could be used throughout NEB rainfall season, even for the least wet months. To make full use of these levels of skill, the model forecasts should be made in standardized units and then converted to observed rainfall using the relationship between the observed standardized and the actual rainfall. In particular, a tendency to overestimate rainfall needs to be corrected.

Figures 7b and 7c directly compare simultaneous 3- and 6-month correlations between the model rainfall and the Hastenrath data for 1948–97 with those between the empirical regression simulations, and hindcasts as in Fig. 5b, and the Hastenrath data. Consistent with Fig. 7a, the model is clearly better in simulation mode than are the empirical simulations, with higher correlations through most of the year. In the dry season the empirical model is not strictly valid and gives near zero correlations. Figures 7b and c also show that HadAM2b is clearly better in simulation mode than the empirical method run in forecast mode as well.

Figure 8 gives a wider-scale picture of the skill of the model in simulating seasonal rainfall in March–May 1947–96 on the model grid scale. This period was chosen as 1997 rainfall data were not available for many
areas. The model grid scale is used to show up local gradients, though a larger space scale would be used to maximize predictability for practical use (our NEB region has an effective resolution of $5^\circ \times 7.5^\circ$). NEB is probably the most skillfully simulated land region in March–May. High skill extends over the Nobre region (see Fig. 1 for definition), but falls off rapidly beyond this area, presumably due to the larger relative impact of internal atmospheric variations (Rowell 1998). Thus the eastern coastal region of Brazil from Cape San Roque southward has much lower skill, as does the southermost part of the Nordeste. Interestingly, relatively high correlations are seen over parts of southern Africa and the southwest United States. Some regions where rainfall is most affected by El Niño (e.g., Ropelewski and Halpert 1987) do not show particularly high skill, as this is the season of generally lowest ENSO-related simulation skill because of the sometimes rapidly changing SST anomalies. But this is the season of highest ENSO impact on the tropical Atlantic (e.g., Sutton et al. 2000). So the high skill in NEB in March–May, the season of the “ENSO predictability barrier,” is all the more striking. Compared to the results for HadAM1 shown in PFJS, HadAM2b shows rather more skill in many parts of the smaller region shown here due to a combination of improved simulations and verification data.

c. Time series of simulated Hastenrath series for March–May

HadAM2b simulations for 1912–98 are plotted against the observed Hastenrath rainfall series in Fig. 9. The ensemble mean March–May rainfall has a highly significant correlation with the observed rainfall ($r = 0.78$), though the model rainfall has a wet bias and a lower standard deviation (1951–80 average = 521 mm, standard deviation = 116 mm) when compared to the Hastenrath series (1951–80 average = 444 mm, standard deviation = 155 mm). The correlation is almost as high in early years, $r(1912–47) = 0.78$, as it is in the more recent years, $r(1948–98) = 0.81$. The variance of the six ensemble members is only a little more than that of the ensemble mean. An analysis of variance test (Rowell et al. 1995; Rowell 1998) was carried out to determine the fraction of the total variance of the ensemble, which is common to all members. This common
Table 2. Skill of HadAM2b simulations of Mar–May rainfall, 1912–98, according to simultaneous observed SOI quint category, nonstandardized and standardized.

<table>
<thead>
<tr>
<th>Quint 1 (El Niño)</th>
<th>Quint 2</th>
<th>Quint 3</th>
<th>Quint 4</th>
<th>Quint 5 (La Niña)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEPSOB Nonstandardized</td>
<td>0.65</td>
<td>0.71</td>
<td>0.94</td>
<td>0.62</td>
</tr>
<tr>
<td>LEPSOB Standardized</td>
<td>0.21</td>
<td>0.30</td>
<td>0.56</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Variance is explained by the only common factor for the six members, the forcing from SST variability, and is 76% of the total variance in model March–May rainfall over 1912–98. The remaining variance is due to random internal atmospheric variations. Since the model can never be expected to simulate this component of variability in phase with that observed, this limits the correlation skill that can be obtained when the model is compared with observations. The formula of Rowell (1998) shows here that this correlation skill limit would be 0.85. So we can conclude that the small shortfall from perfect skill in simulation mode seen in this region is rather more due to random internal atmospheric variations than to errors in responses to SST forcing. Model forecasts would be expected to show a lower skill since the precise patterns of SST anomalies are not known in advance.

AGCMs have sometimes been found to have highest skill during ENSO events (Graham et al. 2000; Brankovic and Palmer 2000), when the SST forcing imposed on the atmosphere may be particularly large relative to internal variations. To investigate this for NEB, the 87 March–May seasons 1912–98 were stratified according to quintls of the March–May SOI calculated using the Trenberth (1984) method. Table 2 shows the correlation and LEPSOB skill for each category. Simulation skill assessed this way without standardization is highest when the SOI is in its middle quint and lowest when the SOI is in the lowest quint, the quint almost invariably associated with El Niño events. When standardization is carried out, this bias disappears. So it is important to standardize the model data. Otherwise the combination of wet bias in the model, and reduced standard deviation compared to the observations, results in poor simulations (and clearly forecasts) in strong El Niño conditions as these are quite often very dry.

Spectral analyses confirm the high skill of the AGCM simulations of March–May rainfall for 1912–98. Sperber and Hameed (1993) showed that March–April 1950–85 NEB rainfall was indeed largely phase locked to a combination of SST variations in the ENSO region in the tropical Pacific and to the tropical Atlantic north–south SST gradient. Both the simulated rainfall and the observed rainfall time series show spectral peaks on the ENSO timescale (around 3–5 yr) and around 11–15 yr.

Figure 10 shows that the squared coherency between these series is above the 95% confidence level [using the method of Bloomfield (1976)] on timescales around 2 and 5 yr and 11–15 yr. Statistical simulations from the WF predictors produced in jackknife mode also show spectral peaks near 3–5 and around 11–15 yr. However, the statistical simulations overestimate variability at 11–15 yr and squared coherence at this timescale is lower than for the AGCM simulations, with less near 2 yr, though a significant peak is still seen on quasi-biennial timescales.

d. Correlation of model simulations with atmospheric circulation

A thorough analysis of the observed variations of NEB rainfall and atmospheric circulation can be found in Hastenrath and Greischar (1993a). Here we show a few averaged results from the ensemble of six HadAM2b runs. Figure 11a shows correlations between HadAM2b rainfall over the four Hastenrath grid points and 925-hPa atmospheric circulation. There are very strong local correlations between enhanced rainfall and enhanced westerly ($r > 0.8$) and northerly components ($r > 0.6$) of the wind flow. This translates into enhanced wind convergence ($r > 0.6$) and moisture convergence ($r > 0.6$) over the NEB coast, and on the southern side of the Atlantic ITCZ. There is enhanced divergence on the northern side of the ITCZ and reduced divergence.

![Squared coherency with observed Hastenrath March-May rainfall](image)
(increased convergence) on the southern side, representing a southward shift in wet conditions. Over the tropical Pacific, moderately enhanced divergence can be seen \((r > 0.4)\), especially over the central Pacific, reflecting the moderately positive correlation with La Niña widely noted above. It is noticeable that the relationships with local 925-hPa meridional wind and divergence are reversed locally just inland over NEB. The reason is not clear, though may indicate a heat low effect similar to that seen with variations of Sahel rainfall (Rowell et al. 1992).

Figure 11b shows comparable results at 200 hPa. The 200-hPa zonal wind over NEB is negatively correlated with rainfall, with easterly anomalies associated with more rainfall. In wet conditions, southerly winds overlie the surface northerly anomalies over and to the north of the north coast. Wet conditions produce increased 200-hPa divergence over NEB and over a more southerly position of the ITCZ in the tropical Atlantic with reduced divergence on the northern side. These results are generally consistent with previous research and show that the strong relationship between the model’s atmospheric circulation variations near NEB and modelled nordeast (Hastenrath grid points) rainfall are dynamically consistent with the strong model SST–rainfall relationships.

e. Hindcasts from the PROVOST experiments using persisted SST

A set of self-consistent dynamical model hindcasts is available from an extension to the European PROVOST...
project, also using HadAM2b. Ensembles of nine hindcasts were made for each season from 1982–93 starting from analysed initial atmospheric conditions (Graham et al. 2000). For the March–June hindcasts the observed February SST anomaly field [using the Reynolds and Smith (1994), optimally interpolated SST data] was persisted throughout the hindcast and added to a climatologically changing field. Figure 12 compares the ensemble mean hindcasts in the peak rainfall season, March–May, using standardized rainfall values based on the AGCM climatology for 1951–80 with observed NEB rainfall data. The latter is calculated as an observed average at the four Hastenrath region model grid points using the Hastenrath data. We show standardized hindcasts to remove model biases as discussed above. To put this period into context, the HadAM2b simulations from 1912–98 show that 1982–93 was a particularly predictable period, with a correlation of 0.95 with the observed data and a LEPS skill score of 0.63. This might explain the remarkable skill of the persisted SST hindcasts that had a correlation of 0.96 with the Hastenrath data. The corresponding LEPS skill is also exceptionally high at 0.80. Thus this period appears not to be typical and to overestimate long-term real-time forecast skill.

7. Real-time forecasts

The Met Office issued 12 sets of real-time forecasts of seasonal NEB rainfall between 1987 and 1998. The three main predictands were: (i) the average rainfall in Fortaleza and Quixeramobim, averaged over March–May (FQ-MM), (ii) the Hastenrath series, averaged over March–April rainfall (H-MA), (iii) Nobre’s February–May rainfall index (N-FM). The reasons for these choices of predicted periods are mainly historical. A series of rainfall averages for Fortaleza and Quixeramobim was available from the Met Office CLIMAT archive for the first forecast issued in 1987. The 2-month Hastenrath and 4-month Nobre series were available for the second issued forecast in 1988. The three series do provide somewhat more information than can be deduced from one series as they cover different periods of the rainy season and somewhat different areas. However, the series are strongly correlated. Two sets of forecasts are made each year, a preliminary forecast around 10 February, and an updated forecast around 10 March. The preliminary forecasts for QF and the Hastenrath region can be regarded as “long lead” because they are issued more than two weeks before the start of their rainfall seasons. However, the preliminary Nobre forecast is regarded as a “nowcast” as the Nobre data includes February. All the updated forecasts are nowcasts. The issued forecasts contain information about the individual prediction schemes, best-estimate quint category forecasts based on that information, and a statement on confidence.

WF gives full details of the two statistical methods used, linear discriminant analysis, and linear regression. Tests of the latter were discussed extensively in earlier sections. Linear discriminant analysis is a key component of the forecasting system as it intrinsically provides probability forecasts that are arguably best matched to the nature of the seasonal forecasting (or any forecasting) problem. In all the issued forecasts described here, linear regression was used in noninflated mode. Judgment of the forecasting team was also used in arriving at the final forecast, such as assessing how the SST patterns were changing at the time of the forecast using half monthly or even pentad SST information. Since 1994, information from AGCM ensemble forecasts based on persisted SST has been included. This has influenced the issued forecast subjectively (most notably in 1997, see below), though only limited information was sometimes available on the model’s rainfall climatology, and more than one version of the AGCM has been used. Thus the way in which information from the various models was combined was not optimal.

The issued forecasts are assessed in section 7a. The component statistical forecasts are briefly assessed in sections 7b and c. The real-time AGCM forecasts are also briefly described in section 7a but cannot be assessed rigorously. They were issued for different regions, involved two different models, and the forecast rainfall anomalies were not based on consistent underlying statistics.

a. Skill of issued long lead and updated forecasts

When the forecast methods are combined, the issued best-estimate forecast is made in one of five rainfall categories each representing an equiprobable 20% (quint) of the cumulative probability distribution of the rainfall totals for 1951–80. Sometimes a forecast is made “on” a quint boundary, reflecting the judgment of the forecasters that the adjacent quints are about equally likely. The very dry quint is referred to as quint 1 (Q1), the dry quint as quint 2 etc., and the very wet quint is quint 5. The observed Nobre data have been refined compared to those described in Carson (1998) and Ropelewski and Folland (1999). The Nobre index was originally updated from 1981 using data for the (smaller) Hastenrath area for the months February–May. Our analyses are now based on the Hulme (1994) data for the complete Nobre area to 1994, but are still based on February–May Hastenrath data thereafter. This is because of lack of recent data in the western part of the Nobre region (see Fig. 1).

Figures 13a–c show the performance of the real-time forecasts in the form of forecast quints plotted against those observed, and are slightly modified from figures in Carson (1998) and Ropelewski and Folland (1999). The modifications included the correction of errors found in our records of the preliminary forecasts for 1988 and 1990. A reassessment of the quint boundaries has also changed the observed quint for FQ in 1996 from quint 5 to quint 4. Because of strong disagreement
between the statistical (dry or very dry) and three AGCM forecasts (wet or very wet) for the preliminary forecast in 1997, a nil best estimate forecast was issued. In Carson (1998) and Ropelewski and Folland (1999) linear regression forecasts are attributed to the issued forecast. Here, we have accepted that an overall nil forecast was issued, but have used the real-time long lead statistical and dynamical forecasts for 1997 when individually assessing them. Table 3 shows the correlation between observed and forecast quint, LEPS skill, bias, and rmse (last two measured in quints) for the six sets of real-time forecasts.

Figure 13d shows a tentative assessment of the potential predictability from the real-time AGCM forecasts alone. Real-time AGCM predictions were made of March–May mean rainfall for the four right-hand grid boxes in Fig. 1 from 1994 onward. The categorical forecasts shown in Fig. 13d were not part of the original issued forecasts, but were calculated later and corrected for the observed bias in the mean and standard deviation of simulated rainfall over the climatological period, 1951–80. At the time the forecasts were issued, the AGCM predictions were only available as anomalies relative to a model climatology that was generally appreciably different from the 1951–80 climatology used for the statistical forecasts. Five years is too short a period to provide statistically reliable skill estimates. Figure 13d shows that the real-time performance of the AGCM has been mixed, with modest skill for the preliminary forecasts, but much better results for the updated forecasts, comparable in skill to the simulations.

The number of issued forecasts varies between 10 and 12, depending on area and the timing of the forecast. Some of the variation in skill between the areas is random. We have summarized the skill using the overall correlation and LEPS scores for (a) long lead forecasts (preliminary FQ and preliminary Hastenrath) and (b) nowcasts (the remaining forecasts). The 21 long lead forecasts defined this way have a correlation of 0.59 between the observed and forecast quint and a LEPS skill of 47%. The 44 nowcasts have a correlation of 0.72 and a LEPS skill of 48%. This arguably indicates a clearly “useful” level of overall skill in both types of forecasts, and is thought to be comparable with the best skill obtained so far over about a decade in real-time seasonal forecasting. These assessments are also consistent with our assessments of the potential predictability of NEB rainy season rainfall.

We now discuss the distribution of forecast and observed quint in more detail for long lead forecasts and nowcasts. Where a forecast was made on a quint boundary, the forecast was allocated to categories on either side of the boundary, with a weight of 0.5 in each category. The full assessment tables are available from the authors. During the period, there was an observed bias to both very wet or very dry conditions compared to the 1951–80 reference period. Combining the three rainfall series together, remembering that only FQ contributes in 1987, the number of observed seasons in quints 1–5 respectively, were 10, 3, 7, 6, and 8, respectively (total 34); 6.8 observed seasons are expected by chance in each quint. This contrasts with Sahel rainfall (Folland et al. 1991; Ward et al. 1993), which has been drier than average in the majority of the seasons for which there...
have been real-time forecasts since 1986. For preliminary forecasts and updated forecasts together, the total number of forecasts in each category, and the number observed in brackets were Q1: 4.5 (20); Q2: 19.5 (6); Q3: 7.5 (11); Q4: 17.5 (12); Q5: 16 (16). Thus there was a clear tendency to issue too few forecasts of quint 1 (very dry), consistent with a nonsignificant but noticeable wet bias in the forecasts shown in Table 3.

PFJS show that, for a quint table, positive LEPSCAT scores are obtained in all cases for forecasting the correct quint and for making an error of one quint. Negative scores are obtained for errors of 2, 3, and 4 quints. We have analyzed the number of forecasts having errors of 0, 1, 2, 3, and 4 quints compared to their chance number (in brackets), allowing for the numbers of observations. For the 21 truly long lead forecasts, these are 0 quint error 7 (4.2); 1 quint error 10 (6.4); 2 quints error 1.5 (4.6); 3 quints error 2.5 (3.8); and 4 quints error 0 (2.0). In all cases where LEPSC scores are positive, the number of forecasts exceeds their chance number and where they are negative, they are less than the chance number. For the 44 nowcasts, comparable statistics are 0 quint error: 15.5 (8.8); 1 quint error 22 (12.4); 2 quints error 5 (10.6); 3 quints error 1.5 (7.0); and 4 quints error 0 (5.2). This confirms that the nowcasts have a similar behavior, with more skill than the long lead forecasts when measured in this simple way.

Summarizing, we note that for long lead and nowcasts together, when the expected number of quint errors from 0 to 4 is compared to their chance values, a chi-squared value of 32.7 is obtained. With four degrees of freedom, this is significant far beyond the 99.99% confidence level. The number of independent long lead forecasts is too few to carry out a chi-squared calculation realistically.

**b. Performance of real-time regression forecasts**

Issued real-time regression forecasts for the three NEB regions have been based on SST averaged over periods preceding the forecast period that varied from one year to another. For the preliminary real-time forecasts, SST was averaged over all or part of the preceding November-January. For the updated forecasts, February was often used but sometimes December, January, and February SST were averaged to provide a better estimate of preseason SST. As described by WF, multiple regression forecasts only started in 1989, so these assessments are only for 1989–98.

Assessments of the real-time regression forecasts are shown in Table 4. Real-time forecast skill is compared against the skill of a set of inflated and noninflated regression hindcasts for 1912–86. We compare hindcasts that are always based on January SST with the preliminary forecasts, and hindcasts always based on February SST with the updated forecasts. Mean LEPS scores are
calculated for the best estimate forecast values using continuous and categorical versions of LEPS. The rmse and mean BIAS of the forecasts are expressed as a ratio of the forecast rmse and bias to the standard deviation of the observations.

We first note that differences in skill between preliminary and updated real-time forecasts are reasonably consistent between the skill measures so that the preliminary forecasts are consistently less skillful than the updated forecasts. The correlation skill implies that about 35% of the variance of the observations is explained by the preliminary forecasts, but near 50% by the updated forecasts (when the forecasts are expressed as a best estimate rainfall value). LEPSCONT, which also assesses forecast values is, as expected, systematically less than LEPSCAT, which assesses forecast categories that cover a range of values. Differences in skill between the regions might mainly be ascribed to chance, though on the grounds of areal size Fortaleza and Quixeramobim might in principle be least skillful, and the Nobre region most skillful. A hint of this behavior can be seen in Table 4 for the long hindcast period 1912–86, but is not clear for the decade of real-time forecasts. The bias is similar between regions and the two forecast periods; it is always positive and typically about 20% of observed standard deviations. Note that the bias in the hindcasts is zero, because the jackknife method forces the hindcast values to be equal to the observed values on average. So in this respect the jackknife method can give different results to forecasts in a period independent of the training period and slightly overestimate skill in a completely independent period. The rmse is higher (around 83% of the standard deviation of the observed) for the preliminary forecasts than for the updated forecasts (around 73% of the observed standard deviation), the equivalent respectively of 54 and 62 mm of rainfall over the season on average. Finally, the standard deviations of the forecast values are typically about 70% of the observations, and show no systematic difference between preliminary and updated forecasts.

The skill of the real-time forecasts can be compared to the potential predictability estimated by the noninflated and inflated regression hindcasts for 1912–96. Inflation does not affect correlation skill so this is the same as noninflated correlation. There is no great difference in correlation skill between preliminary forecasts (0.59) and preliminary hindcasts (0.66). For updated real-time forecasts, the mean correlation skill is 0.69 and for hindcasts it is 0.68. This similarity between real-time forecasts and hindcasts suggests that the period 1989–98 as a whole was less predictable than the remarkably predictable PROVOST period of 1982–93. Figure 13 confirms this pattern of behavior in the real-time forecasts; they were indeed very successful over 1989–93 and less successful afterward.

LEPSCONT and LEPSCAT can highlight the benefits to skill, if any, of inflated regression (not used in real-time forecasts in Table 4). LEPSCAT can also provide more appropriate assessments of the categorical form of the forecasts. Concentrating on LEPSCAT, the mean skill for preliminary real-time forecasts was 37% and for updated forecasts was 47%. For the hindcasts, the comparable values for noninflated regression are 56% and 62%. So in contrast to correlation, this suggests that the potential predictability for noninflated regression forecasts issued in quints was greater in the hindcast than during the real-time forecast period (which did not, however, include the successful 1987 and 1988 forecasts, which used discriminant analysis only). A little of this higher potential predictability may be due to the use of the most recent SST month to calculate the hindcast predictors, which Fig. 5 shows to be marginally the best strategy. However, the 1901–80 WF EOFs and the 1951–80 observed quint categories were evaluated over periods overlapping the hindcast period and might slightly exaggerate hindcast skill. The real-time forecasts were made, of course, for an independent period. However, we do not expect this to have more than a minor effect on skill as demonstrated by WF. Turning to the inflated regression hindcast results, the comparable skills are 61% and 63%. This suggests that there is a minor advantage in using inflated regression at the assessed levels of skill for categorical forecasts made in quints. Such a strategy is confirmed by the LEPSCONT skill that assesses the actual hindcast values. On average, LEPSCONT skill is distinctly greater for both preliminary and updated forecasts. For example, 36% for noninflated preliminary hindcasts and 43% for inflated preliminary hindcasts. In addition, preliminary hindcast rmse is much closer to updated hindcast rmse than in the real-time forecast assessments. As expected, rmse is higher for the inflated regression hindcasts than for the noninflated regression hindcasts. So it seems that the difference between preliminary and updated forecasts may be, in part, data dependent.

c. Performance of real-time linear discriminant probability forecasts

We now assess the real-time discriminant analysis probability forecasts. As with the regression forecasts, these were often chosen from several candidate forecasts using SST averaged over different preceding months, so the use of the SST period was not consistent between forecasts. We include use of the LEPSPROB assessment method described in WF. This was not included in PFJS as in principle this can be “played” but this is not considered a problem here. We also use the ranked probability score (Epstein 1969) in the form described by Daan (1984). Finally, we ask the question: if we had chosen the quint with highest forecast probability to be the categorical quint forecast, what would the LEPSCAT skill have been? This is a realistic question as this interpretation is one of those made when putting together the final forecast. The results are shown in Table 5. The first two scores shown in Table 5 assess the whole

<table>
<thead>
<tr>
<th></th>
<th>FQ</th>
<th>HASTENRATH</th>
<th>NOBRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEPSPROB%</td>
<td>Preliminary</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Updated</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>RPS %</td>
<td>Preliminary</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Updated</td>
<td>44</td>
<td>39</td>
</tr>
<tr>
<td>LEPSCAT %</td>
<td>Preliminary</td>
<td>67</td>
<td>58</td>
</tr>
<tr>
<td>Most probable forecast</td>
<td>Updated</td>
<td>85</td>
<td>70</td>
</tr>
</tbody>
</table>

probability forecast. Again the updated forecasts are more skillful than the preliminary forecasts. LEPSPROB shows a similar performance for all three regions but the RPS scores are somewhat lower. RPS scores are also very sensitive to high probability forecasts for the oppositely forecast extreme category to that observed. This occurred in 1994 and 1996 and this sensitivity adversely affected the RPS skill of the Nobre preliminary forecasts giving a skill score of 0.07.

An outstanding feature of Table 5, however, is the LEPSCAT skills. This has been used to assess the ability of the discriminant analysis method to pick out the most probable category that is regarded as the best estimate forecast. Used this way, discriminant analysis clearly achieved higher real-time skill than did the regression method in Table 4, which, as explained above, was used in “noninflated mode.” Viewed this way, the discriminant analysis forecasts have probably made the largest contribution to the overall forecast skill, though it must be kept in mind that only these forecasts were made for 1987 and 1988, both of which were very skillful. LEPSCAT gives a strikingly different result from RPS for the preliminary Nobre forecasts. This is because the quint forecast with highest probability was observed seven times from 11 preliminary Nobre forecasts. Furthermore, the seven correctly forecast quints were all extreme (very dry or very wet). This gave very high LEPSCAT score of 0.71. LEPSCAT of course has no sensitivity to high forecast probabilities of other categories as long as these were lower, so skill was not offset by this factor.

Besides its intrinsic ability to make probability forecasts, in principle discriminant analysis does not suffer from the reduction in variance suffered by noninflated regression or the somewhat arbitrary results than can occur when regression forecasts are inflated. We can enquire into the performance of discriminant analysis more closely by enquiring how well were the real-time discriminant analysis forecasts calibrated against the observations. To achieve this, we follow the approach of Gilman (1986). Figure 14a shows a plot of the average forecast probabilities for a given quint, binned into increments of 0.25 in the cumulative probability distribution, against the observed frequency of the same quint. To provide sufficient results in each bin, the preliminary and updated forecasts for the three regions (FQ, Hastenrath and Nobre) are assessed together. Quints 2–4 are binned together because of the relatively low numbers of observations of some central quints. Over the period 1987–98, 24 discriminant analysis forecasts were issued for FQ, and 22 for both Hastenrath and Nobre making a total of 68 forecasts (including 1997 when an overall preliminary forecast was not issued).

Ideally the lines plotted for each type of quint forecast should follow the diagonal black line. This joins points for which the average binned forecast probability for a given quint category, and the fraction of such forecasts for which it was observed, are the same. These are the forecasts that are “perfectly calibrated.” Thus for forecasts of quint 1, the point at the bottom left-hand corner...
gives 0.03 as the observed probability of occurrence and 0.08 the forecast probability of occurrence. The value 0.08 represents the average forecast probability for all quint 1 category forecasts made with probabilities in the range 0–0.25. The observed probability represents the fact that for the 39 such forecasts, quint 1 was in fact observed only once. So for this bin, quint 1 forecasts were well calibrated. There are too few examples of forecast probabilities in the range 0.75–1.00 to plot the results for any quint category.

Averaged over all the forecast quint categories, calibration is good. However, low sample sizes may be the reason why the point for forecast quint 2–4 shows a rather poor calibration, though the other two points are well calibrated. For quint 1, there are too many observations of quint 1 for forecast probabilities of quint 1 averaging near 0.4. Instead of observing a 40% likelihood, we observe over 70%. For forecasts of quint 5, we observe too few quint 5 outcomes for forecast probabilities near 0.4. The other two points are well calibrated. Overall, forecasts of quint 1 are above the perfectly calibrated line; this implies that more observations of quint 1 actually occurred than the probability forecasts indicated. Similarly, there were fewer observations of quint 5 than the probability forecasts indicated. This accords with the overall moderate wet bias of the real-time forecasts, which results as much from discriminant analysis as from regression. The fact that the high and low probability forecasts for all quint 1 are well calibrated strongly suggests that there is little tendency for discriminant analysis to produce exaggerated forecasts of high or low probability as once was feared. Thus linear discriminant analysis is a quite powerful statistical forecasting method when used wisely, quite apart from the fact that it is well tuned to the intrinsically probabilistic character of the season forecasting problem.

The sample size is too small to determine whether some of the poorer results in Fig. 14a represent genuine deficiencies of the discriminant analysis equations. So jackknife discriminant analysis hindcasts of the FQ, Hastenrath and Nobre series were produced for 1912–86 in the same way as for the regression hindcasts in Fig. 5. Preliminary hindcasts were based on January SST and updated hindcasts on February SST (Fig. 14b). It can be seen that the hindcasts are well calibrated with forecast probability being close to observed probability. Good calibration does not of itself necessarily reflect good skill; it is a necessary but insufficient condition. However, the evidence from Figs. 14a and b indicates that the linear discriminant forecasting system has good skill; the scope for better calibration (as opposed to somewhat better skill) is small.

d. Failure of the 1998 preliminary (long lead) forecast

This forecast deserves a special mention as this real-time long lead forecast was the least successful, despite being influenced by the strongest mature El Niño of the twentieth century. The presumption that such a strong El Niño would result in a straightforward, skillful forecast, was not borne out (e.g., Nicholls 1999). The observed rainfall was well into the very dry (quint 1) category, at around half the 1951–80 average value. This continues a pattern of extreme droughts observed concurrently with the strong El Niño events of 1958 and 1983. The long lead 1998 forecast (Colman et al. 1998) was made more difficult because of conflicting SST signals from the Atlantic and Pacific Oceans. The strong El Niño event favored a dry season but positive SST anomalies in the Atlantic between 0° and 20°S, and negative anomalies farther southwest, favored above average rainfall. The HadAM2b AGCM (in the same form used for PROVOST) was used to provide the dynamical forecast. This model had been used to produce real-time seasonal forecasts for NEB since 1996. For the updated forecast issued in March, the dynamical forecast was made using the more recent HadAM3 model, which has now replaced HadAM2b.

For the preliminary forecast, using SST up to January, linear regression predicted March–May NEB rainfall to be in the average or wet category. The discriminant analysis forecast was bimodal, however, with above chance probabilities for both dry (quint 2) and wet (quint 4) categories. The nine individual HadAM2b ensemble members all predicted above-average model rainfall, giving an average of 142% of the model climatology. Allowing for model bias, and the error in the model standard deviation over 1951–80, this converts to a quint 5 forecast. A best estimate forecast of above average rainfall was issued largely because of these dynamical forecasts. However, it was noted that HadAM2b predicted below average rainfall over a large area to the east and west of NEB, and also to the north over the ocean. By the time of the updated forecast, the SST signal was more clearly favoring dry conditions due to warming of SST in the northern tropical Atlantic relative to average (a fairly common development in the last stages of an El Niño). The empirical forecasts became drier in response, the regression model favoring quint 2. Updated dynamical forecasts from a nine-member ensemble of the new HadAM3 model were very different from the preliminary HadAM2b forecasts. The HadAM3 ensemble mean forecast rainfall was 36% of the HadAM3 1951–80 climatology, firmly in the quint 1 category.

Forecasters were aware during the preliminary forecast of the possible sensitivity of NEB rainfall to strong ENSO events. Because of the small amount of evidence available (only two ENSO events in the last 50 yr were associated with extreme droughts), they decided it was better to believe the dynamical model forecast of very wet conditions. To test if the HadAM2b model was intrinsically at fault, or whether the observed SST changes between January and the forecast period played a strong role, HadAM2b was rerun with simultaneous SST for
early 1998, extending the long runs of section 6. In contrast to the real-time forecast, this rerun forecast had no input from real atmospheric data prior to the rainfall season. However, the use of simultaneous SST data decreased model rainfall over NEB in March–May only to about the 1951–80 average, though it was simulated to be considerably drier than the average over other parts of Brazil. This suggests that the SST changes between January and the forecast period would have made HadAM2b somewhat drier if they had been known, but by an insufficient amount to rescue the forecast. So the large and beneficial change in the dynamical forecast between the preliminary and the updated forecasts was partly related to the change of model. There are still no long-term assessments of the accuracy of HadAM3 during El Niño conditions in NEB. These are needed to determine if this result reflects a more general improvement during strong ENSOs, or whether internal atmospheric variability not related to SST played a bigger role in 1998 than is usual for NEB rainfall.

8. Discussion: Possibilities for improved forecasts

The real-time NEB rainfall forecasts have relatively high skill when compared with the current state of the art in tropical seasonal forecasting (Hastenrath 1995b). However, there is some evidence that the influence of strong El Niño events on rainfall is underestimated in the empirical methods, indicating that a linear approach is not optimal. A brief investigation was carried out as follows. Hindcast skill scores were calculated over 1948–97 for the regression method using the WF predictors in a slightly modified way. Hastenrath March–May rainfall was hindcast in each case from January SST. First, prior to calculating the regression coefficients, all negative (La Niña) values of the Pacific SST EOF time coefficient were raised to powers of between 0.1 and 5.0 and the hindcast equations recalculated in jackknife mode. No evidence of an increase in skill was found. The same procedure was repeated for positive Pacific SST EOF coefficients (El Niño years). This time, skill modestly increased to reach a flat maximum when the EOF coefficient was raised to the power 1.5 and fell beyond that value. Note the Atlantic EOF coefficient adjusts at the same time as the Pacific coefficient is raised to a power. Over the whole period 1912–97, little increase in skill was seen but over 1948–97 the correlation increased from 0.69 to 0.74 and the LEPSCAT skill increased from 42% to 47%, and the hindcasts correctly on average became drier in strong El Niño conditions. This reflects the approach of Mullan (1995) for New Zealand temperature. He used a bilinear regression between a measure of ENSO and New Zealand temperature that gave greater weight to El Niño than to La Niña for the same “strength” of either. The reality of this nonlinearity needs further investigation, but this modification is being used in forecasts starting from 1999. A further possible influence comes from decadal modulations of El Niño SST patterns, for example, those that have been termed the Interdecadal Pacific Oscillation (IPO, Power et al. 1999). These have been shown to significantly affect El Niño teleconnections to Australian rainfall on seasonal timescales. A way forward is to carry out idealized SST experiments with an AGCM to investigate the joint influences of ENSO and the IPO on tropical rainfall generally, including NEB. Such experiments are underway (B. Mullan 1999, personal communication). The IPO may mainly reflect the average stochastic modulation of El Niño, but the concept provides a framework for looking at this problem.

The experience of the 1998 forecast also underlines the need for the most recent available and reliable SST data. It seems likely that preliminary forecasts issued in early February would benefit from a sufficiently reliable picture of the SST in the last week or so of January. There is still a need to track changes in the SST data before the forecast, and the projected influences of these changes might be integrated more formally when the components of the issued forecast are put together.

It is clear that when used consistently, state-of-the-art climate models should have a strong and increasing role in the forecasts. However, the period of the model rainfall climatology must be consistent with that used by the empirical methods, (now changed to 1961–90), and appropriate model grid points averaged. Using standardized model rainfall, as discussed above, can accommodate biases in the model climatology. Probability forecasts can be based on information from ensemble forecasts and the performance of hindcasts. It could be beneficial to combine empirical and AGCM forecasts optimally, but preliminary studies have shown this to have only a negligible effect on NEB prediction skill. We anticipate that in the long run coupled dynamical models (CGCMs) may provide the best forecasts for NEB: it would be desirable to use these, but only after assessing their intrinsic skill, as the CGCM skill may not match that achieved by existing methods for some time. A promising way forward in the short term is the hybrid use of a CGCM to forecast the El Niño SST signal, which is then applied to an AGCM with persisted SST anomalies elsewhere in a so-called two-tier approach (Ji et al. 1994). A CGCM might also be able to give an indication of El Niño-related tendencies in northern tropical Atlantic SST through the rainfall season, clearly a factor in 1998.

Finally our assessments of potential predictability indicate that useful forecasts of the NEB rainfall season could be made as far ahead as early December, at least for longer averaging periods in the rainy season, using empirical methods. Forecasts at this lead time were introduced in December 1998 for January–June 1999 Hastenrath rainfall, based on November SSTs (Colman and Davey 1998). The current preliminary and updated forecasts are being maintained.
9. Conclusions

We have demonstrated that over the period 1987–98, despite occasional failures, the Met Office real-time seasonal forecasts for NEB rainy season rainfall had quite high skill. The standard correlation between the observed and forecast quints for the overall long lead forecasts was 0.63, and 0.75 for the nowcasts. Measured by the LEPS-CAT skill score, which allows for both bias and errors in the standard deviation of the forecasts (PFJS) and whose numerical values are somewhat lower than for correlation, percentage skill levels were 47% and 48% above chance. The real-time forecast performance is in accord with the extensive evidence given in this paper for the potential predictability of NEB rainfall by both statistical and dynamical methods. When dynamical methods are used, it is important to compensate for biases in the model climatology, mainly for bias and errors in variance. This demands several decades of simulations or hindcasts for the climatological reference period, recently changed to 1961–90, ideally using the same number of ensemble members as used in real-time forecasts.

The relative success of the NEB real-time forecasts resulted from an early recognition of the importance of the joint influences of the gradient of SST across the Tropical Atlantic and the strength and sign of ENSO. To obtain good skill, the relative strengths of these effects must be quantified well. As discussed in WF, allowing for the standard deviations of the SST patterns, on average the tropical Atlantic influence is at least twice that of the tropical Pacific. Our results, however, raise the issue of nonlinear influences of El Niño, though they may not be great. Nonlinearity needs to be distinguished from changes in the El Niño SST patterns themselves, some of which may be interdecadal, and this is an unresolved issue. The most likely influence of such variations is on deep atmospheric convection in the Pacific, modulating NEB rainfall and the nearby Atlantic ITCZ in subtle ways.

The skill of the real-time forecasts is consistent with the extensive tests we have made since 1912 of empirical and dynamical predictability, though the period of the PROVOST experiments, 1982–93 was unusually predictable and gives an overoptimistic indication of potential skill. An intriguing feature of our predictability tests was the finding that SST measured just before the main rainfall season gave empirical skill rather greater than simulations using simultaneous SST. The empirical skill appears to be at least comparable to that using dynamical methods for the preliminary forecasts, though dynamical forecasts may be very competitive once the SST pattern becomes clear. However, existing empirical methods cannot reflect all aspects of SST patterns that might sometimes be important. Nevertheless, these results indicate that empirical methods will be useful for some time into the future in this region, a conclusion in accord with a recent review for some extratropical regions (Anderson et al. 1999).

Finally the assessments of the real-time forecasts, and particularly the discriminant analysis forecasts, question the use of quints because the central quint has no skill. On the other hand, three equiprobable categories (terces) may provide insufficient discrimination for the levels of skill demonstrated here. If it is thought desirable to retain an “average” category, a set of five unequal categories, covering the broadest range at the center of the rainfall distribution, could be investigated using hindcast methods. An optimized combination of probability forecasts made in this mode from the discriminant analysis method, and from an ensemble of dynamical forecasts, may provide a good way forward.

The boreal spring NEB rainy season may be one of the most predictable for any land area (Hastenrath 1995a). This is fortunate: NEB is an important region for seasonal prediction because of its high population and frequent droughts that can seriously disrupt society (Hastenrath 1990). We have shown that a scientific and practical basis now exists for seasonal forecasting agencies to develop more detailed seasonal rainfall forecasts tuned to the needs of society in northeast Brazil.

Acknowledgments. We wish to thank David Sexton and Mark Rodwell for running the HadAM2b simulations, Kelvyn Robertson, Mike Harrison, and Tony Evans for running and providing the AGCM forecasts and PROVOST data, and Mike Hulme (University of East Anglia) for constructing the gridded rainfall data set with support from the U.K. Department of the Environment, Transport, and the Regions. The advice of Iracema Cavalcanti, CPTEC, Brazil is also gratefully acknowledged. This work was supported by the U.K. Ministry of Defence Public Meteorological Service Contract. The PROVOST project (ENV4-CT95-0109) was partially supported by the European Commission.

REFERENCES

Colman, A. W., and M. Davey, 1998: Statistical prediction of January–June 1999 rainfall in Northeast Brazil using input from


Walker, Sir G. T., 1928: Ceara (Brazil) famines and the general air
56, 359–364.
in the north Nordeste of Brazil using eigenvectors of sea-surface
——, ——, K. Maskell, A. W. Colman, D. P. Rowell, and K. B. Lane,
1993: Experimental seasonal forecasting of tropical rainfall at the U.K. Meteorological Office. Prediction of Interannual Clime
——, S. Brooks, and C. K. Folland, 1988: Predictability of seasonal
rainfall in the N Nordeste region of Brazil. Recent Climatic
Change—A Regional Approach, S. Gregory, Ed., Belhaven