EVALUATION OF THE IRI'S "NET ASSESSMENT" SEASONAL CLIMATE FORECASTS 1997–2001

BY L. GODDARD, A. G. BARNSTON, AND S. J. MASON

Overall, the IRI seasonal climate forecasts outperform the atmospheric general circulation models and ENSO-based empirical predictions on which they are based.

Predictions of seasonal climate anomalies have been made in some form for centuries. One of the earliest scientifically based schemes is that of India's Meteorological Department, which has been issuing predictions for the all-India monsoon rainfall since the late nineteenth century using various statistical methods (Blanford 1884; Walker 1923; Bhalme et al. 1986). One of the more recent developments is the regular production of seasonal climate forecasts that are based, at least in part, on global dynamical climate models [e.g., the U.K. Met Office (UKMO) since 1988, Ward et al. 1993; the Climate Prediction Center (CPC) since 1994, O'Lenic 1994; the Canadian Meteorological Centre (CMC) since 1995, Servranckx et al. 1999; Derome et al. 2001; Australia's Bureau of Meteorology (BoM) since 1997, Frederiksen et al. 2001].

The International Research Institute for Climate Prediction (IRI) began issuing quarterly seasonal forecasts of global climate in October 1997 (Mason et al. 1999). These forecasts are the subjective assessment and consolidation of many climate prediction tools, and are thus called "net assessments." They provided probabilistic forecasts for below-, near-, and above-normal precipitation and temperature for the upcoming 3-month period and the subsequent 3-month period (more information available online at http://iri.columbia.edu/climate/forecast/net_asmt/). Since mid-2001, however, the IRI has issued these forecasts monthly for the upcoming four overlapping seasons.

Validation studies of the prediction tools used by various meteorological centers suggest that they can provide potentially useful information on seasonal climate variability for many parts of the world. For many regions physical dynamical models replicate, and in some cases improve upon, the skill realized from empirical statistical models, and this has given greater credibility to seasonal predictions obtained by both approaches. Thus, regional managers in sectors such as agriculture, energy, and water resources have

AFFILIATIONS: GODDARD, BARNSTON, AND MASON—International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York
CORRESPONDING AUTHOR: L. Goddard, International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Rt. 9W, Palisades, NY 10964
E-mail: goddard@iri.columbia.edu
DOI: 10.1175/BAMS-84-12-1761

In final form 25 May 2003

recently begun to factor seasonal climate forecast information into planning decisions [Golnaraghi 1997; the National Oceanic and Atmospheric Administration (NOAA) 1999; Agrawala et al. 2001]. An impediment to the use of many operational forecasts, such as those issued by IRI, has been lack of knowledge about the real-time operational skill of the forecasts. Skill measures exist for the constituent prediction tools, but because the IRI net assessment forecasts are subjectively constructed, unbiased attempts to forecast previous years retrospectively are not possible. Now that the net assessments have a verifiable history of more than 4 yr, from October–December (OND) 1997–2001 (hereafter 3-month periods will be denoted by the first letter of each month respectively), their performance can be examined.

A diagnostic verification of the net assessments was recently performed by Wilks and Godfrey (2002, hereafter WG2002) for the period 1997–2000. This diagnostic verification examines the full joint frequency distribution of the forecasts and the corresponding observations. The results of WG2002 yielded useful information about the biases and reliability[1] of the IRI net assessment forecasts. For example, they found that the temperature forecasts issued for 1998–2000 contained a strong cold bias by not predicting the degree to which above-normal temperatures dominated most land areas throughout this period. The temperature forecasts were also determined to be too confident for all three categories at both low and high latitudes. On the other hand, the forecast probabilities for precipitation during 1997–2000 were found to be reliable. The diagnostic verification analysis of WG2002 requires many forecasts in order to examine reliability; they consider all forecast seasons together, and average statistics over space as well as over time. Thus, maps showing the spatial distribution of the biases and reliability characteristics were not possible for the small set of forecasts available.

The results presented in this study complement those of WG2002. We only use the scalar skill measure of the ranked probability skill score (RPSS; Epstein 1969). This measure can give an overly pessimistic view of the performance of the forecasts (Wilks 2000), but because it is a tough metric, regional skill becomes more noteworthy. Using a scalar metric also allows us to produce maps showing the spatial character of the forecast skill, which provides more useful information to those using the forecasts regionally. From maps of RPSS one can examine the spatial distribution of the skill presented in an overall form in WG2002.

Of particular interest to forecasters, as well as forecast users who may be contemplating the use of a particular prediction methodology, is to know whether the subjective human element is adding to forecast skill, and if so, to what extent. In this paper the skill of the net assessments is compared to that of the primary input tools, namely, the dynamical climate model forecasts and a relatively simple empirical prediction method based on ENSO phases. The results provide guidance to IRI forecasters regarding their use of, and confidence in, individual prediction tools. This analysis also provides important regional guidance to current and potential users of the IRI net assessments.

IRI SEASONAL CLIMATE FORECASTS. The IRI began issuing forecasts for 3-month total precipitation in October 1997, and for 3-month mean temperature in January 1998. Currently the IRI issues two types of forecasts. The first, "three category" forecasts, gives the probabilities that seasonally averaged precipitation and temperature will be above-, near-, and below-normal. The three categories are defined from 30 yr of historical data, such that each of the categories is equiprobable. A forecast of "climatology" implies no information beyond the historically expected 33.3%–33.3%–33.3% probabilities.[2] The second type of forecasts, that of extremes, is a more qualitative indication of enhanced risk for the seasonally averaged temperature and precipitation occurring in the upper or lower 15th percentile of the 30-yr historical distribution. For example, the forecast may indicate a 25%–40% probability of extremely above-normal rainfall, which we define as an approximate doubling of risk over the climatologically expected 15% probability. This validation study concerns only the three-category forecasts.

Although the majority of the constituent prediction tools for the IRI net assessments are objective, the

[1] Reliability refers to the correspondence between the probabilities given in the forecasts and the subsequent relative frequencies of occurrence in the observations. For reliable forecasts, these two quantities are equal (see Wilks 1995; Murphy 1993, 1997).

[2] Where no region is explicitly designated, a forecast for climatological probabilities is implied. For precipitation, explicit forecasts are also not issued for "dry regions," which are indicated. Over these areas the climatological rainfall is so low (less than 30 mm over the 3-month season) that the below-normal category cannot be well distinguished from the near-normal category.

subjective element of human intervention is not negligible as they are combined to arrive at the final forecast product. Recent improvements to the IRI forecast system have greatly reduced the amount of subjective work necessary to produce the net assessments (Barnston et al. 2003). More objective approaches, such as multimodel ensembling (Rajagopalan et al. 2002), now consolidate much of the information that had been done subjectively. However, the human element is still fundamental to the final forecast product.

Constituent prediction tools. Among the tools subjectively considered and synthesized in the IRI net assessments, atmospheric global climate models (AGCMs) are the most heavily weighted of the prediction tools considered.[3] Where the AGCM(s) have skill historically (see appendix A) and are predicting probabilities different from climatology, the prediction is considered for inclusion in the final forecast. The IRI uses several AGCMs that produce seasonal climate predictions as input to the final IRI seasonal forecast. During 1997–2001, three AGCMs were used regularly: version 3.2 of the Community Climate Model (CCM3.2), developed at the National Centers for Atmospheric Research (NCAR; Hack et al. 1998) and run at the IRI; ECHAM3.6, developed by the Max-Planck Institut für Meteorologie [Deutsche Klimarechenzentrum (DKRZ) 1992] and run at the IRI; and the Medium-Range Forecast model (MRF9) developed by NCEP Environmental Prediction (Livezey et al. 1996) and run by our collaborators at the Queensland Centre for Climate Applications in Australia.

At this time, no observed atmospheric or land surface conditions are used to initialize the AGCM predictions. The initial conditions for the seasonal predictions are taken from ongoing updates to long-run simulations that have been forced with observed SSTs. These initial conditions constitute the models' view of the current atmospheric state together with a reasonable degree of uncertainty, intended to represent the range of plausible atmospheric states consistent with prescribed surface boundary conditions.

The AGCMs are typically forced with more than one scenario of predicted sea surface temperature anomalies (SSTAs) enabling diagnosis of climate sensitivities to differences in SST predictions. One scenario assumes the SST anomalies will persist through the forecast season. These persisted SSTA predictions are used only for the 1-month lead forecasts. The other scenario uses predictions of evolving SSTA in the global Tropics and damped persistence of observed SSTA in the midlatitudes. This second scenario, which we refer to as "forecast SSTA," employs different methods of SSTA prediction in each tropical ocean basin (see appendix B).

The IRI also uses empirical data for forecast guidance. Since ENSO affects climate worldwide, although only robustly over about 25% of land areas (Mason and Goddard 2001), the probabilities of below-, near-, and above-normal temperature and precipitation conditioned on ENSO regularly factor into the net assessments, particularly during El Niño and La Niña events. This tool has been modified to look at not only the strongest warm or cold events but also the "nearest neighbors" (Lall and Sharma 1996) to the forecast ENSO SST index. This tool is based on the NINO3.4 index (the SSTA area averaged over 5°S–5°N, 170°–120°W).[4] We identify the 10 yr since 1950 for which the observed NINO3.4 value (in the analogous season) was closest to that predicted for the upcoming season. The climate anomalies for those 10 historical seasons are then categorized and tallied yielding the ENSO-associated probabilistic prediction.

In addition to the objective prediction tools, we have often incorporated the official forecasts issued by national meteorological services where and when available. If the IRI has no strong evidence that contradicts the national meteorological service, we defer to their forecast. The national meteorological services whose forecasts have been available to us, at least occasionally, include those of Australia, Brazil, Canada, India, New Zealand, the Philippines, South Africa, and the United States. Some regions, such as southeastern South America, southern Africa, and the Greater Horn of Africa, hold regular regional climate outlook forums, during which regional and international climate forecasters meet to develop a consensus probabilistic climate outlook for the upcoming 3–4-month season (Basher et al. 2001). These regional

[3] Note that a subtle distinction is drawn between predictions, which are considered the objective output of particular empirical or dynamical tools, and forecasts, which may be objective or subjective but that constitute the best available guidance offered to users by the forecaster.

[4] During 1997–2001, the net assessments used the coupled model predictions from NCEP for the tropical Pacific SST (Ji et al. 1998). The ENSO-associated probabilities, as well as the AGCM predictions, are based on this product. During 1997–2000, the seasonally averaged observed and predicted Niño-3.4 SST index agree well at both 1- and 4-month lead times (Fig. 1).

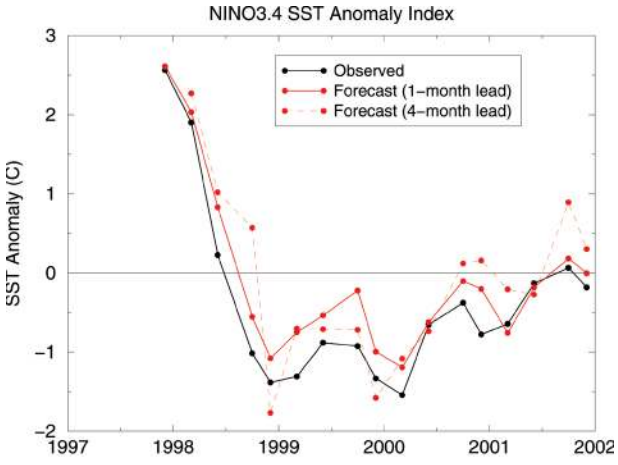

FIG. 1. Time series of the 3-month-averaged NINO3.4 index (°C) (SST anomaly area-averaged over 5°S–5°N; 170°–120°W) from observations (black line; Reynolds and Smith 1994) and from predictions made by the NCEP coupled ocean–atmosphere model (Ji et al. 1998) at 1- (solid red line) and 4-month lead (dashed red line).

forecasts are also considered in the net assessments, when available.

Verification data. TEMPERATURE. The Climate Anomaly Monitoring System (CAMS) dataset (Ropelewski et al. 1985) serves as verification data for near-surface air temperature. These monthly averaged data are based on over 1200 gauge observations that were applied to a 2° × 2° lat–long global grid using the Cressman scheme (Cressman 1959; M. Halpert 2002, personal communication); missing data points are identified. Sea surface temperature values are used over ocean points. This dataset supplies temperature anomalies relative to the 1971–2000 climatological base period. We further removed the 30-yr residual mean bias of the base period appropriate to the forecast: 1961–90 for forecasts issued through 2000, and 1969–98 for forecasts issued after 2000. In order to define the anomalies these data require an adequate station time series, and unfortunately some regions are poorly sampled. Thus, large areas over South America, Africa, and Asia are masked in the analyses due to data missing more than 10% of the time.

Precipitation. Precipitation verification data comes from CPC Merged Analysis of Precipitation (CMAP) dataset (Xie and Arkin 1997). Both rain gauges and satellite observations supply input to the monthly averaged product, which is available globally at 2.5° × 2.5° resolution. There are no missing data points. Because CMAP only covers 1979 to the present, we obtained the climatology and terciles for 1961–90 from the precipitation data of the University of East Anglia's Climate Research Unit (New et al. 1999, 2000). Although it is not ideal to use two separate datasets for the precipitation, it is unavoidable, since no dataset for precipitation is currently available in near–real time that extends back far enough to define the historical climatological characteristics. No large biases were found between the two datasets in the period of their overlap.

VERIFICATION MEASURE. For meteorologists, a clear distinction exists between accuracy and reliability. Perfect reliability and perfect accuracy are synonymous only when the observed category is predicted every time with 100% confidence. Because of the inherent uncertainty in the climate system due to chaos in the atmospheric dynamics and limitations in specifying initial conditions, seasonal climate forecasts cannot be given with 100% confidence. Thus, the goal of climate forecasters is to assign reliable probabilities to categorical forecasts such that the forecast probability for a particular outcome is consistent with the observed frequency of that outcome over time.

For users of climate forecasts, there is great temptation to focus on the accuracy of the forecasts. Too often verification of probabilistic forecasts is incorrectly oversimplified by interpreting the probabilistic forecasts deterministically (taking the category with the greatest probability as *the* forecast category) and comparing them to observations. However, judging a particular forecast in this way, as "right" or "wrong," ignores the information provided regarding the inherent uncertainty for the outcome of a particular category.

Probabilistic forecasts should be scored according to a measure that appropriately treats their probabilistic information, such as the RPSS. See appendix C for derivation and examples of the RPSS. The RPSS gives more credit for forecasting the observed category with high probabilities; however, the penalties for forecasting the wrong category with high probabilities are substantial. The maximum RPSS is 1, but the score will be expressed in this analysis as a percentage of maximum. A score of 100% could only be obtained by forecasting the observed category with a 100% probability consistently. A score of 0 implies no skill in the forecasts, which is the same score one would get by consistently issuing a forecast of climatology. A negative score suggests that the forecasts are underperforming climatology. For typical climate forecasts in which skill is generally modest and in which forecast probabilities therefore typically fall within 20% of their climatological values (of 33.3%),

RPSS scores are often in the range of 5%–20%. To give a point of reference for those not familiar with RPSS, an RPSS of 10% for a three-category forecast system approximately equals a correlation coefficient of 0.50 (correlating the median of the probability distribution).

As noted by Wilks (2000), the RPSS scalar measure of performance often presents an overly pessimistic view of forecast performance. Many contributions to the performance are represented as a single score, such as variability of the median of the probability distribution as well as of the spread of the distribution. Because errors in any of the contributions adversely affect the score, the RPSS is a stringent test of forecast performance. Furthermore, different forecasts with very different error characteristics could score identically, and examination of the scalar score alone would not permit the differences to be diagnosed (Wilks 2000). For these reasons, the more complete diagnostic verification of WG2002 should be considered together with the results presented in this analysis, which specifically targets the comparison of skill between the net assessments and the constituent objective prediction tools.

RESULTS. *Overall skill of net assessments.* In evaluating the net assessments using the RPSS, one may choose to assess only those times and locations for which a deliberate forecast was made (Figs. 2a, 3a), or one may choose to assess all points at all times, including the climatological forecasts (Figs. 2b, 3b). Comparison of Figs. 2a and 3a shows that on average higher skill is obtained when the climatological forecasts are excluded, implying that for regions containing nonclimatological forecasts the category having the highest forecast probability was observed more often than one-third of the time (i.e., that expected a priori). Note that in Figs. 2a and 3a, local skill scores may be based on a reduced sample of forecasts, and thus may be subject to high sampling variability. At some high-latitude locations, for example, fewer than four nonclimatology forecasts may have been issued.

Positive skill exists over a majority of the land area for both the temperature and precipitation net assess-

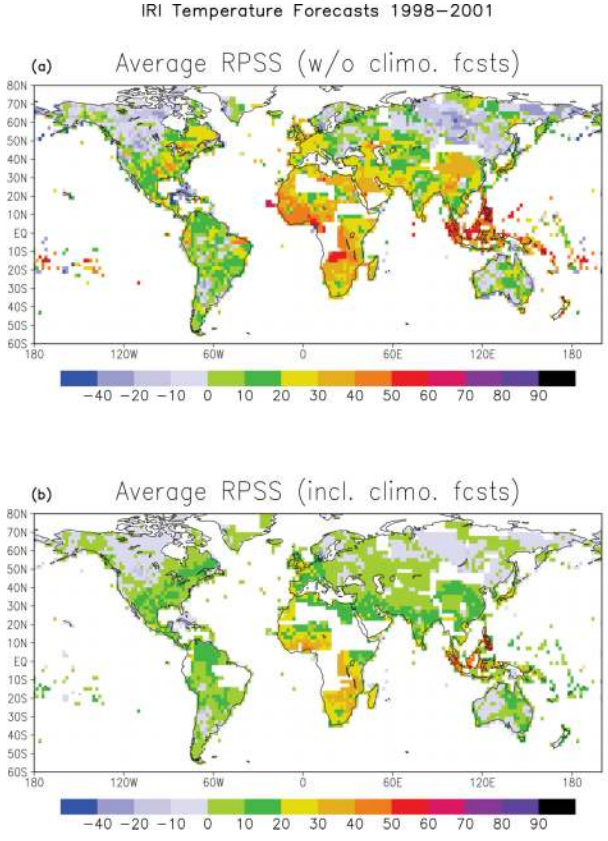

FIG. 2. Geographical distribution of RPSS (%) averaged over 16 quarterly IRI net assessment forecasts of near-surface air temperature, JFM 1998–OND 2001. (a) Avg score calculated excluding climatology forecasts. (b) Avg score calculated including climatology forecasts. Blank areas over land indicate grid points for which observed data record covers less than 90% of period.

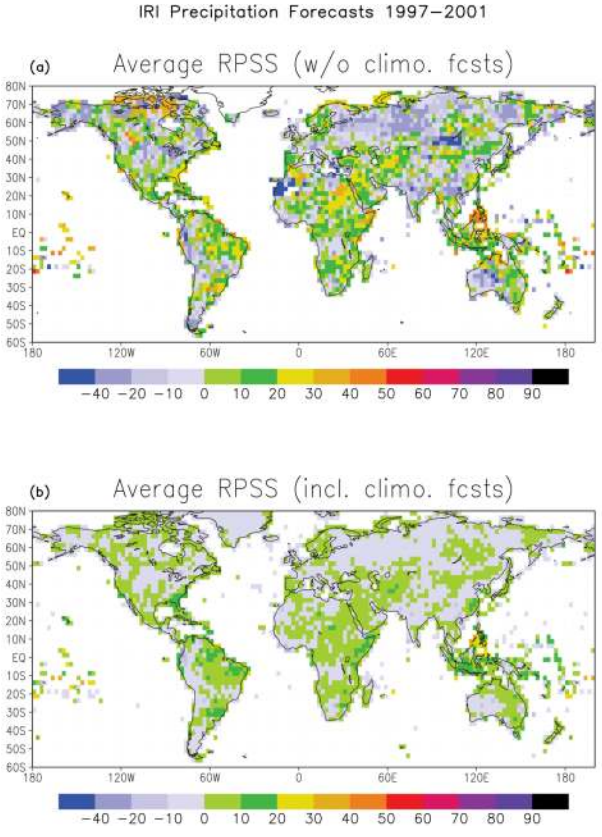

FIG. 3. Same as in Fig. 2, but for 17 quarterly IRI net assessment forecasts of precipitation: OND 1997–2001.

ment forecasts. As seen in most predictability studies, temperature variability is predicted with higher skill and better coverage than is precipitation variability (e.g., Mason et al. 1999; Peng et al. 2000; Derome et al. 2001). For temperature, higher skill is found over the Tropics, but even midlatitude regions, such as western Europe and eastern Canada, show relatively high levels of skill for 1998–2001 (Fig. 2). For precipitation, higher skill over the Tropics is not as marked, but many regions documented as potentially predictable (Hastenrath 1995) appear with positive skill here, such as the Indonesia region, the tropical Pacific islands, eastern Africa, northern South America, and the southeastern United States.

The temporal evolution of net assessment forecast skill over the 4-yr period is shown in Figs. 4 and 5. The averaged skills are positive except for the case of OND 1998. Even at 4-month lead time the skill is consistently positive, and tends to be only slightly lower than that obtained from the 1-month lead forecasts. Overall positive skill is also found at the continental scale (Table 1).

No obvious relationship appears between the strength of ENSO forcing in the tropical Pacific and skill in near-surface air temperature forecasts over land. Higher skill is seen in the RPSS of temperature during mid-1998, which was a time of transition when the El Niño of 1997/98 was giving way to the development of La Niña conditions that persisted through early 2000. The season of lowest skill is clearly OND 1998. At this time the La Niña event was at full

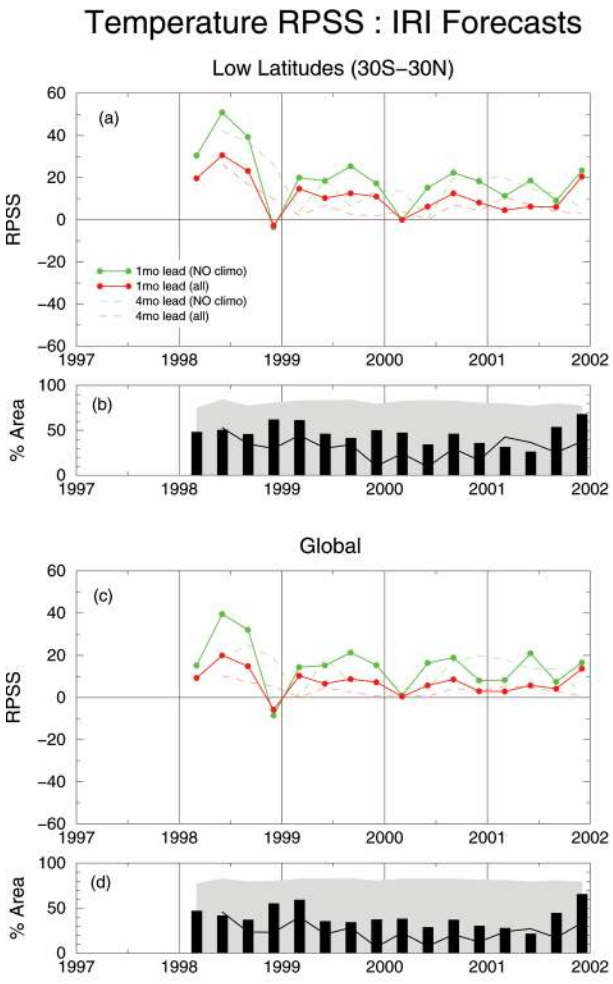

FIG. 4. (a) Time series of RPSS for IRI net assessment temperature forecasts area averaged over tropical land points (25°S–25°N), for 1-month (4-month) lead times shown by solid (dashed) line showing skill for nonclimatology forecasts only in green and for all forecasts including climatology ones in red. (b) Percentage of land area over which nonclimatology forecasts were issued over tropical region for 1- (bars) and 4-month lead forecasts (black line). Gray area indicates percentage of tropical land points over which observed verification data is available. (c), (d) Similar to (a) and (b), respectively, but averaged over global land points (60°S–80°N).

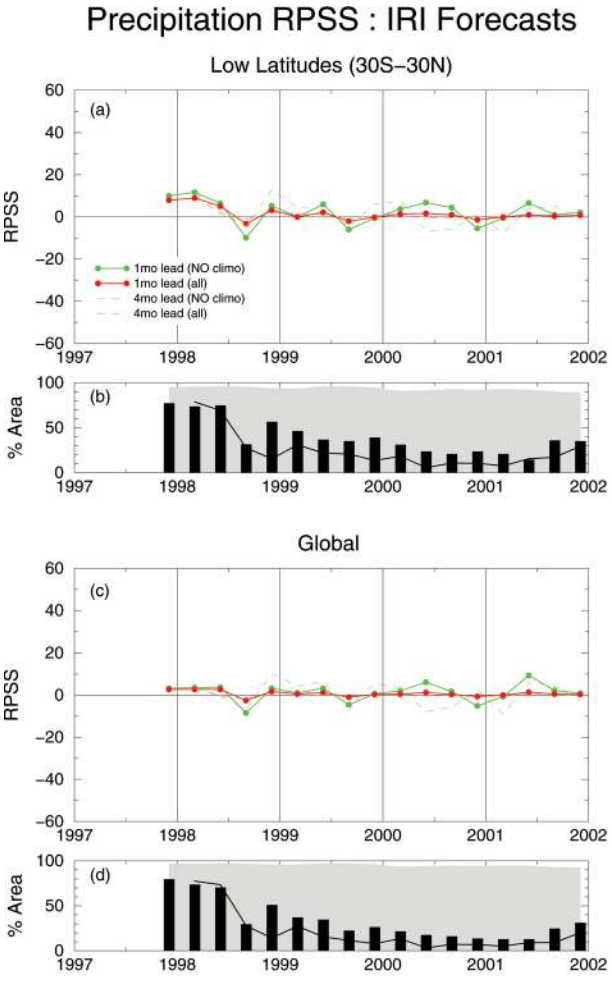

FIG. 5. Same as in Fig. 4, but for precipitation.

strength, and as is shown later, the AGCMs did a good job responding to that forcing. However, the IRI forecast for temperature put too much confidence in historical ENSO-related probabilities, which validated poorly. It was decided after that one season that ENSO-related probabilities would not be a useful tool for temperature predictions during this multiyear La Niña episode, and it was subsequently dropped from consideration for the temperature forecasts.

Area-averaged RPSS for precipitation is modest (Fig. 5). However, skill is positive for most forecasts, particularly in the Tropics, and particularly for the 1-month lead forecast season. The precipitation skill in the Tropics is greater during the El Niño event of 1997/98, but clearly positive skill is again seen in 2000 and 2001, when tropical Pacific SSTs were weak and without coherent structure.

The seemingly modest skill of the precipitation forecasts may just reflect lower potential predictability for precipitation than for temperature. Examination of the forecast reliability (Table 2) indicates that the precipitation forecast probabilities were approximately reliable. The results in Table 2 are consistent with the finding of WG2002: more confident forecasts of above- or below-normal precipitation carry an increased occurrence of that category in the observations; forecasts favoring the near-normal category have no skill and should probably not be issued; a wet bias is indicated in the forecasts. On this final point, it is true that the observations were much more frequently below-normal than above-normal, and that the forecasts did not indicate enhanced probabilities for below-normal precipitation in enough cases. However, the forecasts did indicate more regions over which probabilities favored below-normal than above-normal precipitation, suggesting that the tendency for drier conditions throughout the

TABLE 1. Area-averaged RPSS of IRI net assessment precipitation forecasts (%) for 1-month lead forecasts (4-month lead in parentheses) (i.e., perfect skill = 100%).

	JFM	AMJ	JAS	OND
Globe	5.7 (2.1)	9.5 (5.0)	9.0 (4.3)	4.5 (2.5)
Tropics (20°S–20°N)	8.5 (6.1)	10.3 (6.3)	14.0 (9.0)	13.3 (5.9)
Africa (40°S–40°N, 30°W–60°E)	14.2 (1.7)	14.9 (9.6)	20.0 (8.8)	22.4 (8.3)
Asia (10°S–80°N, 55°E–170°W)	4.1 (6.1)	7.6 (3.2)	7.2 (5.1)	7.4 (1.7)
Australia (40°–10°S, 110°–155°E)	5.7 (8.5)	12.4 (8.2)	116 (4.7)	9.2 (3.4)
Europe (30°–85°N, 30°W–65°E)	2.8 (2.2)	9.0 (0.5)	5.0 (2.8)	4.7 (–0.1)
North America (0°–85°N, 170°–45°W)	3.5 (–4.7)	6.8 (3.5)	4.7 (3.2)	–6.3 (0.9)
South America (60°–15°N, 95°–35°W)	10.5 (6.0)	13.4 (12.4)	8.9 (3.8)	–12.1 (2.0)

TABLE 2. Tables showing the relative frequency with which the observed category verified given the particular categorical forecast issued at the stated confidence level. Forecast categories are listed across rows and observed categories are listed down columns. The values in parentheses show the total number (over grid points and forecast occurrences) that a given case occurred. For example, the upper-left box in each table represents the fraction (number) of times that above-normal precipitation was observed when a forecast was issued for above-normal seasonal precipitation. Perfect reliability would appear as diagonal elements, from upper left to lower right, equal to the stated confidence level.

40% Confidence:	A_0	N_0	B_0
A_f	**0.35**(487)	0.23(317)	0.42(578)
N_f	0.26(267)	**0.26**(268)	0.49(506)
B_f	0.26(507)	0.23(446)	**0.52**(1028)

45% Confidence:	A_0	N_0	B_0
A_f	**0.31**(312)	0.29(287)	0.40(398)
N_f	0.36(257)	**0.26**(188)	0.38(276)
B_f	0.24(255)	0.16(166)	**0.60**(620)

≥ 50% Confidence:	A_0	N_0	B_0
A_f	**0.50**(535)	0.27(290)	0.23(241)
N_f	0.95(18)	0.05(1)	0.00(0)
B_f	0.14(140)	0.18(182)	**0.67**(662)

lower latitudes during 1998–2001 was captured in the forecasts.

The authors speculate that the pervasive below-normal precipitation was the result of significant above-normal SSTs in the western Pacific and Indian Ocean regions. Increased SST in regions that are climatologically convective adds greater heating to the atmosphere locally, encouraging more convergence into the region, and thus drawing convective rainfall from tropical land areas and increasing subsidence over land (Graham 1995; Kumar et al. 2003). Furthermore, enhanced heating in the Tropics is likely to strengthen the atmospheric Hadley circulation leading to increased subsidence and drier conditions over the subtropics (e.g., Oort and Yienger 1996). Limitations in AGCM representation of the convective regions of the Tropics, particularly in the spatial structure of the convective regions, may account for the differences in spatial coverage of below-normal precipitation seen in the predictions compared to the observations. Table 2 considers the performance of the precipitation forecasts over low latitudes (30°S–30°N), but similar results for forecast reliability apply to the midlatitudes, although much fewer forecasts were issued with probabilities greater than 50% for the dominant category.

The percentage of land area covered by nonclimatological forecast probabilities appears to show some association with ENSO strength, reaching about 75% during OND 1997. However, the drop in coverage following 1997 has also to do with the evolution of the forecast methodology, including increasing caution in forecasting over the midlatitudes and the introduction of dry region masking (see footnote 2). Furthermore, ENSO exhibits statistically robust teleconnections for precipitation over only 20%–30% of the land in any one season (Mason and Goddard 2001). Thus, in addition to the assumptions of typical ENSO impacts, the extensive areal forecast coverage during OND 1997 and JFM 1998 can be attributed to anomalous SSTs throughout the global Tropics, to which the models responded through shifts in regional seasonal climate probabilities (e.g., Goddard and Graham 1999; Goddard et al. 2001). The anomalous SSTs in the tropical oceans are largely captured in the IRI's predicted SST scenarios that force the global climate models, particularly for the 1-month lead forecasts.

Skill of net assessment versus objective prediction tools. The results shown in Figs. 2–5 and previously described document the overall skill of the IRI net assessment forecasts during the OND 1997–2001 period, both spatially and temporally. However, the value of the net assessments is better discerned by comparing their skill against those of the constituent objective predictions.

Temperature Forecast Comparison. In Fig. 6, the RPSS of the IRI net assessment temperature forecasts are compared against the ENSO-based probability predictions and the AGCMs that served as inputs to the net assessments. (Hereafter, unless otherwise stated, the skill results for the net assessments refer to those including forecasts for climatological probabilities.) The ENSO-based prediction (Fig. 6b) is clearly the tool of lowest skill during this 4-yr period. Many of the areas over which the net assessments showed the highest skill for 1998–2001, and ENSO-associated probabilities showed the lowest skill, including northern South America, tropical Africa, China, and western Indonesia, which have experienced particularly persistent above-normal temperatures during these 4 yr relative to the climatological base period of 1961–90. The ENSO-based tool failed because 1998–2001 was dominated by La Niña, which historically had been associated with predominantly below-normal surface air temperatures over the Tropics.

The AGCM temperature predictions show high skill over many areas (Fig. 6c–6e). Both the AGCM predictions and the net assessments yield extensive coverage of positive skill. These results indicate that even when forced with predicted SSTs, the AGCMs perform well for temperature. Over longer validation periods, in which the AGCMs were forced with simultaneous observed SSTA, many of the same regions with positive skill in this evaluation do show good skill, such as Central and South America, tropical and southern Africa, and the Indonesian region. Over 1998–2001 many other regions showing high skill for temperature, such as eastern Canada, western Europe, and China are not areas of high skill in the historical simulations (of 1965–97). During 1998–2001, the RPSS for the AGCMs are generally higher than those for the net assessments over areas where both are positive.

This suggests that perhaps we should place more confidence in the probabilities determined by the models. However, closer examination of the skill over time (Fig. 7) shows that most of the AGCMs' advantage is due to their very high skill during 1998. Note that the skill of all AGCMs is higher during the transition from El Niño to La Niña during 1998 than it was during the peak of the El Niño (JFM 1998) or the subsequent La Niña years (Fig. 7). This result is not yet well understood; however, it is worth noting that

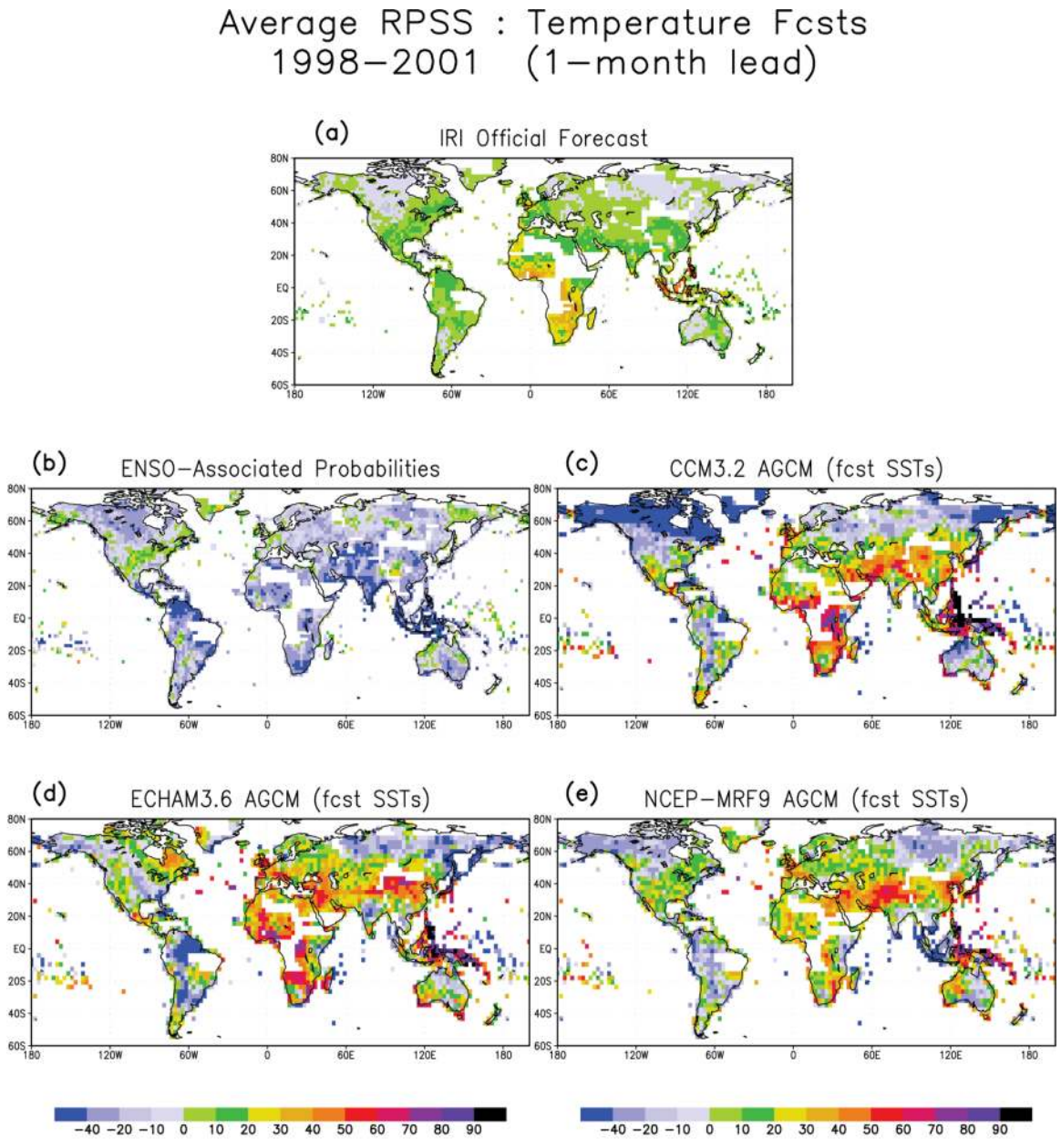

Fig. 6. RPSS (%) averaged over 16 (1998–2001) seasonal forecasts of near-surface air temperature at 1-month lead time from (a) IRI net assessment forecasts, (b) ENSO-associated temperature probabilities, conditioned on predictions of Niño-3.4 from CPC coupled ocean–atmosphere model; (c) CCM3.2 AGCM forced with predicted evolving SSTA scenario; (d) ECHAM3.6 AGCM similarly forced; and (e) NCEP–MRF9 AGCM similarly forced.

although the maximum anomaly in tropical SST was realized in December 1997, the maximum in total tropical SST occurred in April 1998, and the maximum SST anomaly over the convective regions of the tropical oceans (taken approximately as 10°S–10°N, 60°E–180°) reached its peak in July 1998. Since 1998, the net assessment has scored comparably to, and often higher than, the AGCM forecasts.

One may ask whether the good skill for temperature during 1998–2001 resulted from the persistence of warm temperatures over the period. Although the AGCMs did predict a dominance of above-average temperatures and those predictions verified more frequently than did the areas predicted to be below normal, historical simulations do not indicate that the AGCMs are more skillful at predicting above-normal temperature than they are at predicting below-normal temperature. WG2002 found that the net assessment forecasts underpredicted above-normal temperatures and overpredicted below-normal temperatures. This

is also the case for the AGCMs, and to a greater degree than found by WG2002 for the net assessments. The IRI forecasters were aware that cold conditions were being overforecast and corrected the situation somewhat, although clearly not enough. This subjective correction is evident in Fig. 7, particularly in 2000 and 2001 when the AGCMs' skill is lowest.

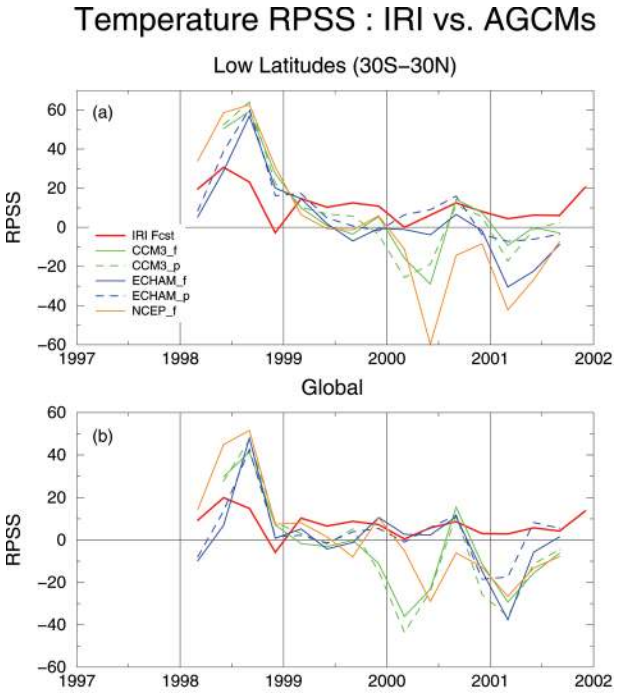

Fig. 7. Time series of RPSS for 1-month lead temperature forecasts area-averaged over (a) tropical (25°S–25°N), and (b) global (60°S–80°N) land points, comparing IRI net assessment forecast (red line), and the CCM3.2 (green line), ECHAM3.6 (blue line), and NCEP–MRF9 (orange line) AGCMs forced with either persisted SSTA scenario ("_p," dashed lines) or the predicted evolving SSTA scenario ("_f," solid lines). The RPSS does not show much difference in skill between the two different SST scenarios, although there is a weak suggestion of higher skill over the Tropics under the persisted SSTA scenario. In the persisted scenario the tropical SST anomalies are often stronger and closer to the observed SSTA during the 1-month lead forecast, relative to the evolving SSTA prediction. Warm anomalies tended to dominate the tropical SSTA pattern during 1998–2001, and those were better represented in the persisted SSTA scenario. The evolving SSTA predictions, which are statistically based in the tropical Atlantic and Indian Oceans, damp the SST toward climatology during seasons for which the SST prediction models do not have significant skill, resulting in weaker warm forcing of the global Tropics. However, regional and area-averaged skill differences were greater among the various AGCMs for a particular SSTA prediction strategy than they were between the two SSTA prediction strategies for a particular AGCM.

Precipitation Forecast Comparison. Precipitation variability for many regions depends strongly on the seasonal cycle. Furthermore, regions with well-defined wet seasons typically have activities such as agriculture dependent on a wet season, which makes forecasts for that season important. For this reason, we discuss the precipitation skill comparisons for several specific regions during their wet season. These regions were chosen because the potential predictability for their wet season has been documented.

January–March (JFM). The west coast of the United States has a well-defined wet season from approximately December to March. The southeastern region of the country, while not subject to the same well-defined rainy season, does experience significant interannual variability during JFM. The skill of the net assessment forecasts (Fig. 8a) is positive across the southern tier of the United States, with highest skill found in the southwest and southeast. The objective prediction tools also show good skill over this region. Over the western/southwestern United States and western Canada, the net assessment forecasts give the best coverage of positive skill, although locally, specific tools have higher skill. Over the southeastern United States, all model predictions (Figs. 8c–8e) show excellent skill, outscoring the net assessment forecasts and ENSO-associated probabilities. Over Mexico, the ENSO-associated probabilities and the models all appear to have some skill; however, over a multidecadal historical record these tools are not skillful over this area in JFM. The AGCMs and the ENSO-associated probabilities do have historical skill (not shown) over the southeastern United States, and some suggestion of skill in the west of the country.

For southern Africa, the main rainy season is December–March when the ITCZ is in its southernmost position. The ENSO-associated probabilities and the CCM3 and ECHAM3 AGCMs show skill for JFM historically, and would be expected to do so during 1998–2001. Positive skill exists for most of the tools over southern Mozambique and Zimbabwe. This is consistent with historical skill analyses of CCM3 and ENSO-associated probabilities. Positive RPSSs for the net assessments cover a broad and coherent region that includes this area and more.

Indonesia and Australia are directly impacted by the shift of western Pacific convection associated with ENSO. During JFM positive RPSSs for historical AGCM simulations exist only over the Philippines.

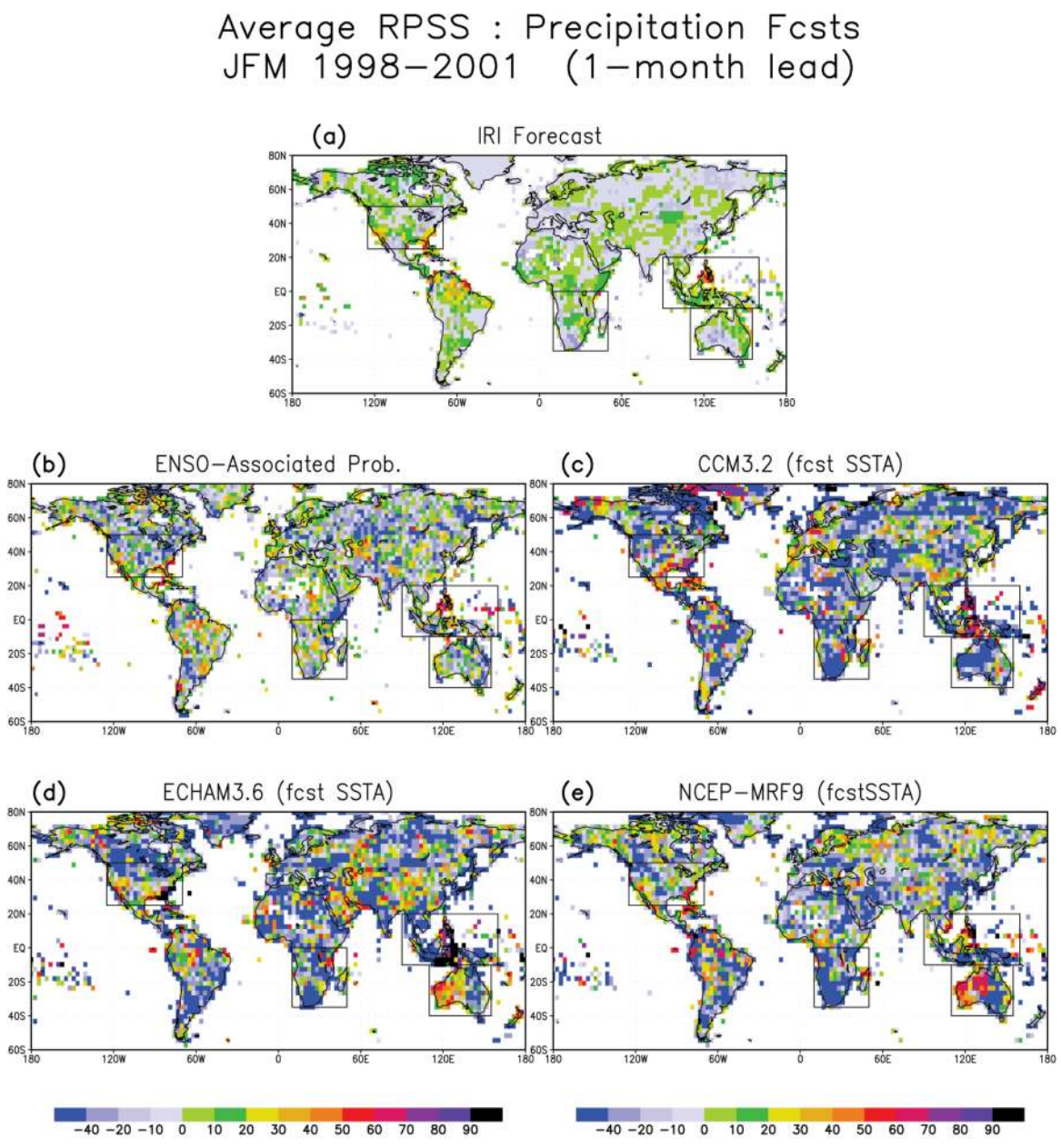

FIG. 8. Same as in Fig. 6, but for JFM precipitation. Score is averaged over four cases (1998–2001). Boxes outline the regions for which skill comparisons are highlighted in text.

The ENSO-associated probabilities show positive skill over central and eastern Indonesia and along the coast of Queensland in northeastern Australia. During JFM for 1998–2001 all objective prediction tools and the net assessment demonstrated good skill over much of Indonesia, and also over the Philippines (Fig. 8). In fact, over the Philippines and over central Indonesia, the AGCMs exhibited higher skill than the ENSO-associated probability predictions. The AGCMs forced with the evolving SSTA scenario show better performance than those forced with the persisted SSTA scenario (not shown). This result is not surprising since tropical Pacific SST anomalies are typically weaker in March, by the end of the forecast period, than they are in November (the month from which the persisted SSTA for a JFM forecast would have come). The ECHAM3 and NCEP AGCMs also performed very well over western Australia (Figs. 8d–8e); however, neither of these AGCMs performed particularly well over this region in simulations of the previous decades. Again, although the objective tools locally have higher skill than the net assessments, the net assessments have more coherent coverage of positive skill.

April–June (AMJ). The region of northeast Brazil has very high predictability during its rainy season (e.g., Hastenrath et al. 1984; Ward and Folland 1991), which covers February–May. Most of the interannual variability for the entire rainy season, however, materializes during April and May. Again, the net assessments show a more coherent region of positive skill than the objective prediction tools during AMJ over northeastern Brazil (Fig. 9a). For this region, the ENSO-associated probabilities scored the lowest overall during AMJ for 1998–2001, even though this region is one of those more routinely influenced by ENSO (Mason and Goddard 2001). Most of the AGCM predictions performed well over parts of northeastern Brazil; however, none of them show positive skill throughout the very important coastal region of the Nordeste. This may be the result of imperfect SST forecasts in the tropical Atlantic. In historical simulations with these AGCMs using observed SSTs, this region has the highest RPSS of any region during AMJ. The AGCM predictions counted heavily in the net assessments and exhibit generally positive

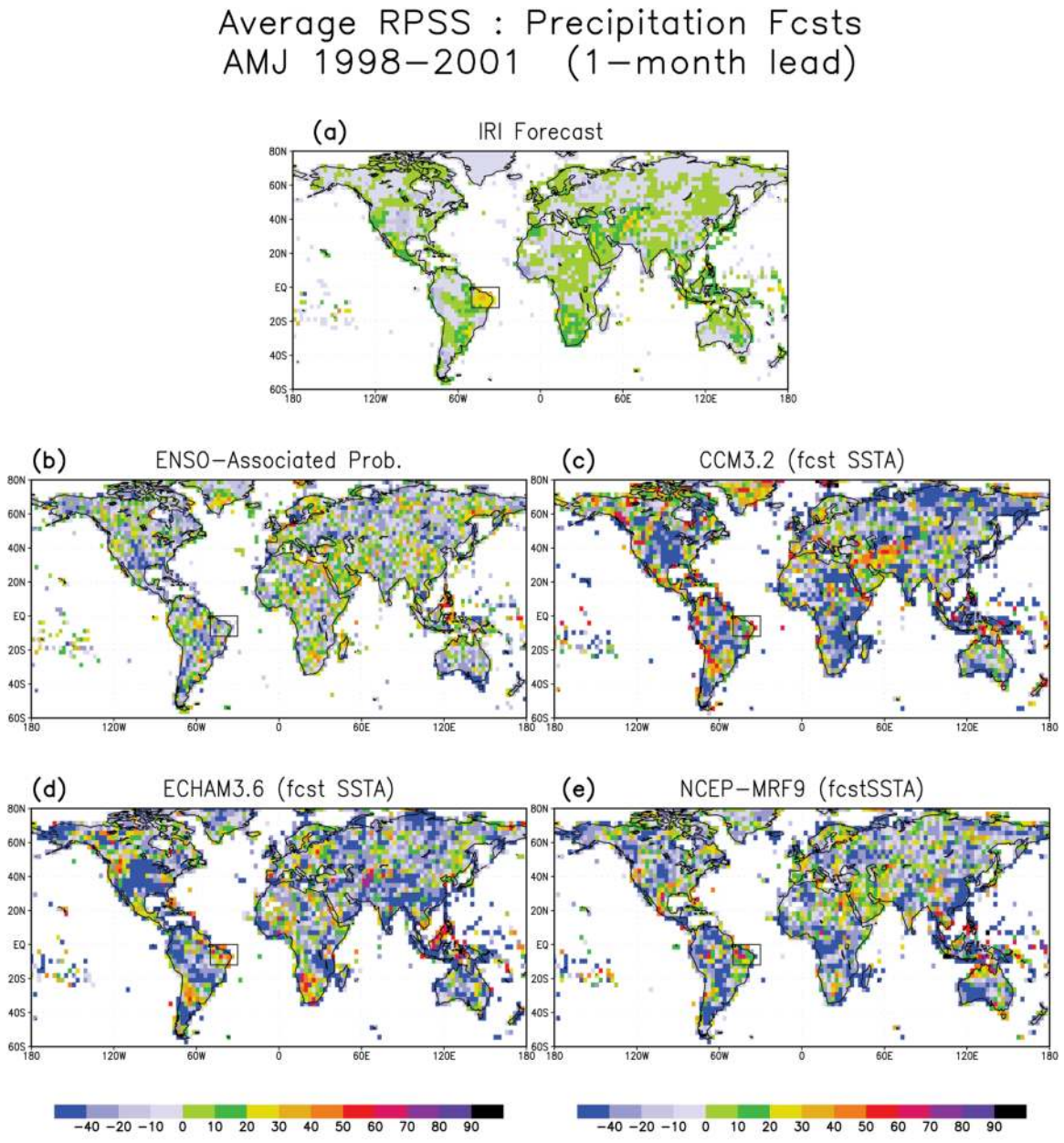

Fig. 9. Same as in Fig. 6, but for AMJ precipitation. Score is averaged over four cases (1998–2001). Boxes outline the regions for which skill comparisons are highlighted in text.

skill; however, it appears that subjective human input contributed to a greater areal extent of positive forecast skill for northeastern Brazil.

July–September (JAS). During JAS, rainfall associated with the intertropical convergence zone (ITCZ) reaches its northernmost extent to the Sahelian region of Africa, making this an important season for rainfall predictions there. Of the existing tools, only the ECHAM3 AGCM has demonstrated positive RPSS over the Sahel in the historical record. During JAS 1998–2001, however, most of the objective predictions show scattered positive skill over the region (Fig. 10). For the entire Sahel region, spanning approximately 10°–17.5°N, 17.5°W–25°E, the net assessments show the best areal coverage of positive skill. Although the objective predictions may show higher skill in localized areas, the irregular and sparse distribution of these areas suggests that the overall structure of the anomalous climate signals is not being captured.

The southwestern Indian monsoon occurs during June–September. The variability in monsoon rains has

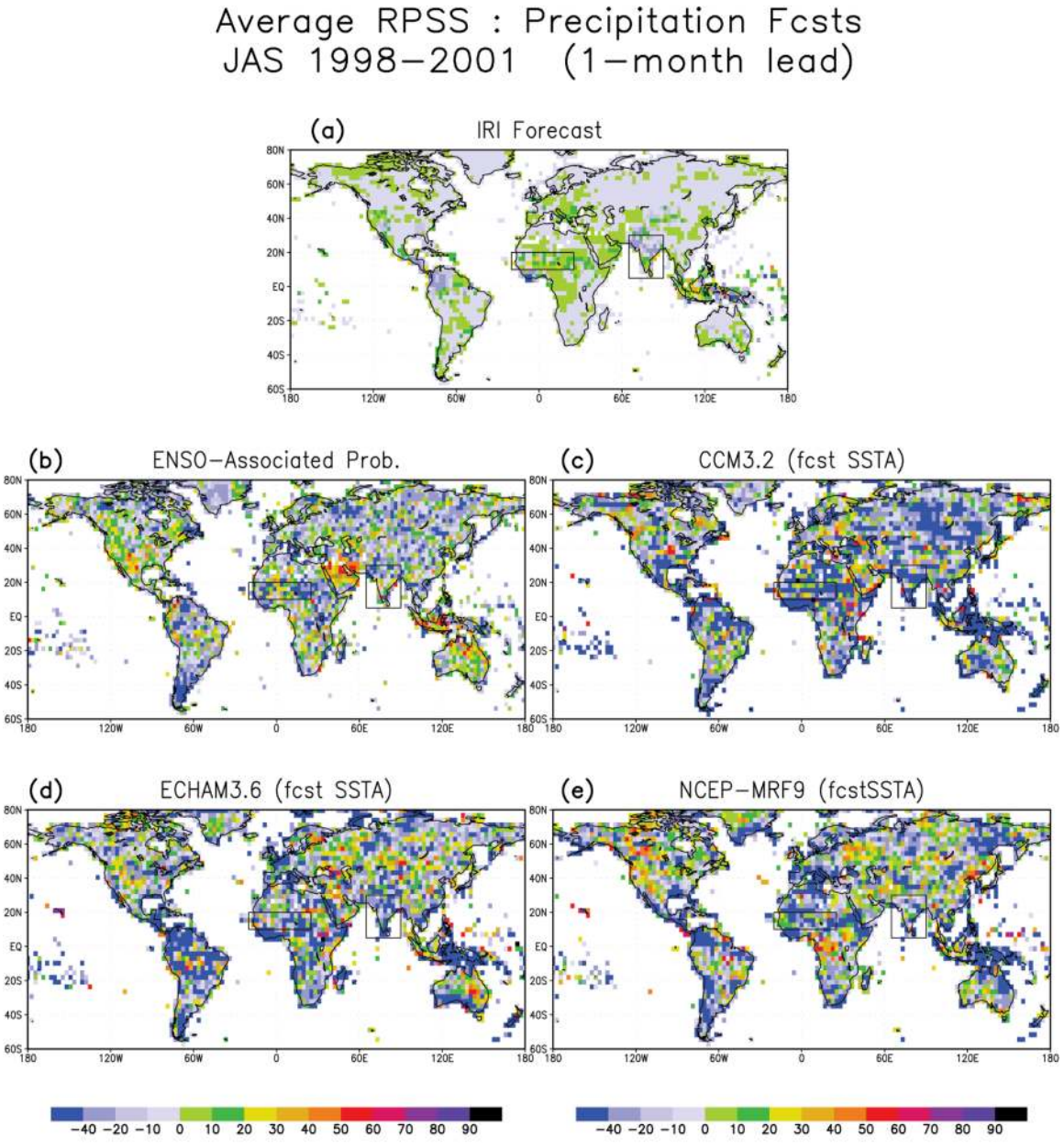

Fig. 10. Same as in Fig. 6, but for JAS precipitation. Score is averaged over four cases (1998–2001). Boxes outline the regions for which skill comparisons are highlighted in text.

been linked to ENSO variability (e.g., Shukla and Paolino 1983) and other dynamical variability over the Indian Ocean region (e.g., Kawamura et al. 2001). Although statistical predictability has been demonstrated, the AGCMs typically do not perform well over the Indian subcontinent during June–September. The ECHAM3 and NCEP AGCMs do exhibit a small region of positive RPSS historically during JAS, but the skill of the ENSO-associated probabilities is higher and more widespread during this season. During the 1998–2001 forecast period, the ENSO-associated probability skill (Fig. 10b) outperforms the AGCMs (Figs. 10c–10e) over southern India. Positive skill for the net assessments does cover southern India more coherently than the best performing objective tool. In this case, the subjective element led to more coherent positive skill by spatially smoothing the prediction indicated by the ENSO-associated probabilities, which reduced the spatial noise of the prediction.

October–December (OND). The rainfall variability of southeastern Brazil has good predictability during OND and experiences a statistically significant association with ENSO events, particularly La Niña events (Mason and Goddard 2001). The ENSO-associated probabilities score better than other objective predictions for southernmost Brazil and northern Uruguay during the OND season 1997–2001 (Fig. 11). The AGCMs do show areas of positive skill, but the positive skill of the net assessments shows a smoother, more coherent pattern. The region of positive skill in the AGCMs and ENSO-associated probabilities over southern Brazil, to the north of Paraguay, does not exist in the IRI forecast. Historical analyses indicate that ENSO-associated probabilities and the CCM3 model both have positive RPSS skill in the area. It appears that these predictions were not given enough credit in the production of the net assessment forecasts.

Over the Greater Horn of Africa the variability of the OND rainfall is significantly associated with ENSO (Farmer 1988; Beltrando and Camberlin 1993; Mutai et al. 1998; Mason and Goddard 2001), albeit indirectly through changes in Indian Ocean SSTs (Goddard and Graham 1999). The historical RPSS of the AGCMs suggests some predictability in this region, but it is not high and does not cover much area. However, other skill measures, such as correlation, indicate quite significant predictability for eastern Africa (not shown); thus, the low RPSS in the historical runs may reflect overconfidence in the AGCMs' probabilities. For the OND 1997–2001 set of forecasts, the AGCMs again show spotty regions of positive skill (Figs. 11c–11e). The ENSO-associated probabilities (Fig. 11b) perform better than the AGCMs, and historically this is true also. The net assessment forecasts (Fig. 11a) again show a larger and more coherent region of positive skill than seen in any of the individual objective predictions.

OND is a time of transition from the east Asian monsoon to the Australian monsoon and also a time when ENSO variability, which directly impacts the climate over the Indonesian region, typically peaks. Over Indonesia and Australia the historical skill, as measured by RPSS, is strong for the ENSO-associated probabilities, averaging at about 20% over the region. For the AGCMs the historical RPSS is weak, and positive values are not coherent across the region. However, during OND 1997–2001, the AGCMs, particularly those forced with evolving SSTA predictions, performed equal to or better than the ENSO-associated probability predictions over central Indonesia and the Philippines (Figs. 11b–11e). The IRI forecast skill is positive and coherent throughout this region and even shows positive skill over Sumatra and New Guinea where there is less consistent skill among the objective tools.

Table 3 summarizes the precipitation skill averaged over the regions described above, and shown in Figs. 8–11, for all seasons. Area-averaged scores are also provided for the continental regions, for the Tropics, and for all global land areas from the IRI net assessments and, in parentheses, the ENSO-associated probability scores. The overall skill is positive, with the exception of JAS. The lower average RPSS skill in JAS precipitation results in part from the difficulty of predicting the variability of the monsoon systems active in that season. At the continental and regional scale the net assessments generally score positively, and they outscore ENSO-associated probabilities in most cases. For the majority of the regions discussed above, the wet season is the season with the highest skill for the 1997–2001 period. The exceptions are southern Africa for which the forecasts performed best on average for the AMJ and OND seasons than for the JFM season, and for the region covering Indonesia, parts of southeastern Asia, and northern Australia (see Fig. 11), for which the JFM forecasts had slightly higher average skill than the OND forecasts.

SUMMARY. The IRI net assessment forecasts cover the period of late 1997 to the present. These probabilistic three-category forecasts for temperature and precipitation result from subjective consideration and synthesis of many objective prediction tools, with considerable weight given to the predictions from

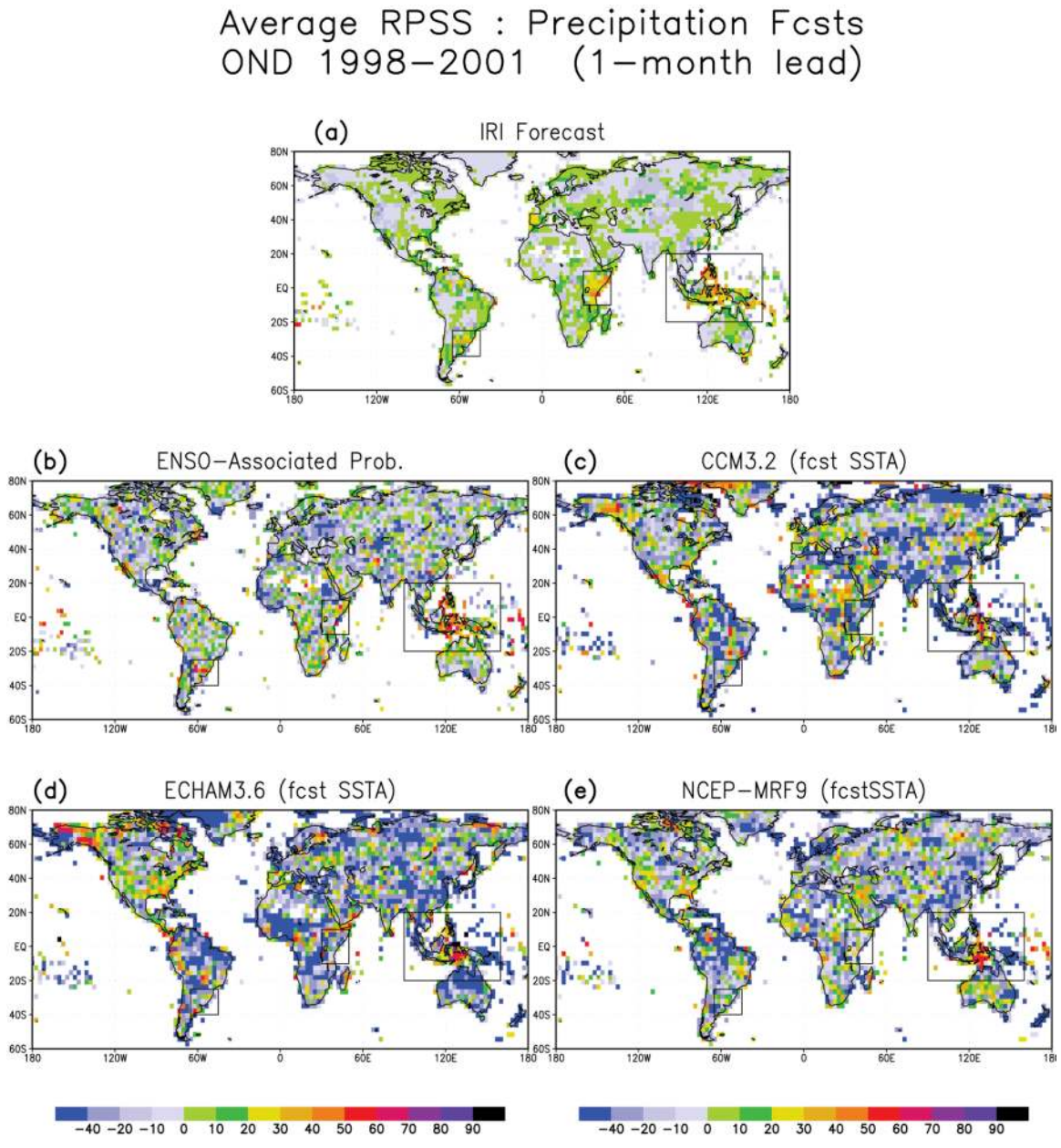

FIG. 11. Same as in Fig. 6, but for OND precipitation. Score is averaged over five cases (1997–2001). Boxes outline the regions for which skill comparisons are highlighted in text.

atmospheric global climate models (AGCMs). Although the IRI forecast system is becoming increasingly objective (Barnston et al. 2003), forecasts issued during the 1997–2001 period relied heavily on forecaster interpretation of the objective predictions.

The quarterly forecasts for the period OND 1997 through OND 2001 are judged in this paper using the ranked probability skill score (RPSS), a measure that, although stringent, does consider the probabilistic content of the forecasts. It is recommended that the results from this analysis should be considered in combination with the more thorough diagnostic verification of Wilks and Godfrey (2002). Nonetheless, they demonstrate that the skill of the net assessments is a clear improvement over the skill that would have been achieved by other means of prediction, such as empirical predictions based on ENSO or that from the individual AGCM predictions. The comparison between the net assessments and the constituent prediction tools illuminates to what degree the subjective element of the net assessments is aiding or impairing the forecast product.

It is difficult to accurately estimate the skill of the various tools and of the net assessments given such a

Table 3. Area-averaged RPSS of IRI net assessment precipitation forecasts (%), and ENSO-associated probability predictions in parentheses. Values in bold highlight the seasonal area-averaged RPSS values of regions for which the performance of the wet season prediction is discussed in the text.

	JFM	AMJ	JAS	OND
Globe	0.9 (–5.8)	1.6 (–5.9)	–0.7 (–9.2)	0.8 (–6.9)
Tropics (20°S–20°N)	3.1 (–3.3)	2.7 (–3.7)	–1.1 (–9.3)	2.6 (–3.3)
Africa (40°S–40°N, 30°W–60°E)	0.1 (–6.0)	3.0 (–1.0)	1.0 (–8.9)	1.9 (–7.5)
GHA	8.0 (–19.5)	1.3 (–3.9)	–0.8 (–15.1)	**13.5 (2.9)**
Sahel	4.2 (–1.8)	0.4 (–0.3)	**7.5 (–21.1)**	–2.8 (–12.2)
South Africa	**–0.9 (–4.3)**	3.2 (–1.3)	–0.2 (–9.4)	4.1 (–0.9)
Asia (10°S–80°N, 55°E–170°W)	1.0 (–7.1)	2.2 (–4.6)	–2.0 (9.3)	0.1 (–6.3)
India	–1.6 (–14.7)	–1.4 (–9.6)	**–6.4 (–11.6)**	–2.7 (–16.7)
Australia (40°–10°S, 110°–155°E)	–0.9 (-8.7)	–0.1 (–11.0)	0.3 (2.6)	–0.9 (–0.9)
Indonesia–Australia	6.7 (–0.6)	5.0 (1.3)	–3.1 (–3.1)	**5.7 (6.2)**
Indonesia	**7.4 (1.2)**	5.4 (–0.1)	–3.6 (–4.7)	6.0 (6.5)
Europe (30°–85°N, 30°W–65°E)	–2.9 (–4.4)	1.7 (–3.3)	–0.1 (–10.9)	1.1 (–12.8)
North America (0°–85°N, 170°–45°W)	2.9 (–6.3)	0.3 (–11.5)	–1.3 (–6.9)	0.8 (–8.3)
United States	**3.6 (–6.5)**	–1.2 (–14.1)	–0.9 (5.9)	–2.2 (–8.9)
South America (60°–15°N, 95°–35°W)	4.5 (–5.0)	2.0 (–9.6)	–1.9 (–16.7)	2.0 (–5.4)
Northeast Brazil	5.5 (4.5)	**18.5 (–11.8)**	–0.6 (–3.9)	3.8 (–3.7)
South America	4.1 (–8.8)	7.2 (–2.2)	0.8 (–25.7)	**7.7 (1.8)**

small sample of verifiable forecasts. The details shown in maps and time series include considerable sampling noise that will not necessarily remain constant in the future. However, some conclusions can be drawn about the net assessments and the constituent prediction tools even over this relatively short period.

Overall, the skill of both the temperature and precipitation net assessments does not seem to be dominated by ENSO, nor does the area covered by nonclimatological forecasts. Forecast skill within the Tropics was somewhat better during the peak of El Niño, particularly for precipitation, but skill is generally positive from late 1997 to 2001 regardless of ENSO phase or strength. This period saw particularly warm temperatures throughout the Tropics and even over much of the midlatitudes, and also a preponderance of below-normal precipitation throughout the lower latitudes. These are features one might associate more with warm conditions in the tropical Pacific rather than cold conditions, although El Niño was only in place for approximately 6 months out of the greater than 4-yr period examined here.

The greatest overall skill for temperature was realized during the transition between El Niño and La Niña, rather than during the peak of either one. Similar increases in globally averaged skill of all the AGCM predictions during the ENSO transition suggest that the potential predictability was high that season and that the dynamical tools were able to capitalize on that predictability. This result requires

further investigation, but it appears to be related to the anomalous state of the Tropics as a whole rather than merely to the state of the eastern equatorial Pacific, a finding that emphasizes the need to move beyond ENSO to consider the broader issue of climate prediction.

During 1997–2001, the overall precipitation skill remained approximately constant globally and tropically, although the area covered by nonclimatology forecasts did not. The decrease in area covered by nonclimatology forecasts for precipitation occurred for reasons other than ENSO variability. First, the forecasting philosophy has evolved since 1997. Changes occurred in SST prediction strategies, in the methods used to estimate the AGCMs' probabilistic predictions, and even in the specific use of some tools. For example, the ENSO-based statistical tool was excluded from input to the temperature forecasts following the verification of the first couple of seasons when cool conditions dominated the tropical Pacific while the tropical land temperatures remained warm (see also Kumar et al. 2001), which was uncommon in the historical record, and for some regions unprecedented. Second, dry season masking was introduced to the precipitation forecasts in mid-1998, coincident with the most noticeable decrease in the percentage of area over which a nonclimatological forecast is issued.

Some of the conclusions reached by the IRI forecasters as a result of their experiences during this period and of the evaluation presented here include the following.

For temperature:

- Net assessment forecasts and AGCM predictions yielded high skill during the 1998–2001 period, but the spatial extent of above-normal temperatures was underforecast. The same conclusion regarding a cold bias in the forecasts was reached in the diagnostics verification of the IRI net assessments (Wilks and Godfrey 2002) and of the CPC long-lead outlooks for the period 1995–98 (Wilks 2000).
- ENSO-associated probabilities did not perform well during this period. La Niña events are normally associated with below-normal temperatures over land, but during the La Niña of 1998–2000 warm temperatures dominated most land areas. This poor performance is likely the result of unprecedented warming over the convectively active regions of the tropical oceans, such as the western Pacific and the Indian Oceans, which enhances the strength of the tropical convection and thus the heating of the troposphere.
- Given the dominance of above-normal temperatures and the relative lack of below-normal temperatures throughout the entire verification period, the use of a 30-yr period for defining climatological normals should be reconsidered. A frequently updated 10-yr normal may be more appropriate and informative for user needs.

For precipitation:

- The net assessment forecasts yield more coherent coverage of positive skill over potentially predictable regions, and more positive skill overall, than from any single prediction tool.
- ENSO-associated probabilities are a useful tool for precipitation overall, but they often do not outperform the AGCMs even in areas with known ENSO teleconnections such as Australia, Indonesia, the southern tier of the United States, and northeastern Brazil. This tool is most useful during the limited times when a moderate-to-strong ENSO event is mature in the tropical Pacific. Compared to the ENSO-based statistical tool, the AGCMs are better able to capture the more subtle forcing of the tropical global oceans that may not be directly related to ENSO, particularly at times when the tropical Pacific forcing is weak.
- Forecasts over regions for which coherent patterns of positive skill exist should be considered areas where the forecasts are potentially useable for input to decision making. In cases where the regions' rainfall variability is seasonal, the forecasts are generally most skillful during the historically defined wet season.

The IRI forecast system is continuously evolving. Insights gained from this analysis, and others such as Goddard and Mason (2002), should aid the subjective element in the short run and add to improvements in the objective system in the long run.

The forecast evaluation presented in this paper is only one way of viewing the performance of the forecasts. We have, however, subjected the forecasts to a very strict measure of forecast performance, and regions that exhibit good skill in this analysis are likely to appear skillful when subject to other verification measures. Ultimately, a meteorologist's determination of a skillful forecast is only valuable to the extent that the forecast can provide benefit to those incorporating the information into their decision process. Thus, a forecast provider's definition of forecast quality may vary greatly from a user's definition of quality, and such a definition may even vary among different users (Hartmann et al. 2002). We welcome others to

analyze the forecasts for themselves. The net assessment forecast data are freely available through the IRI data library (see online at http://iridl.ldeo.columbia.edu/SOURCES/.IRI/.FD/.Seasonal_Forecast).

ACKNOWLEDGMENTS. This paper was funded by a cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA) NA07GP0213. The views expressed herein are those of the authors and do not necessarily reflect the view of NOAA or any of its subagencies.

APPENDIX A: USE OF THE CONSTITUENT ATMOSPHERIC MODELS. Multidecadal simulations of approximately 50 yr, using observed SSTs, have been produced for each of the constituent atmospheric general circulation models (AGCMs), each of which has produced at least nine ensemble members. These long historical runs provide estimates of model potential predictability and characteristics of model climatology that are essential to interpreting the seasonal predictions from each model (Mason et al. 1999). Work is under way to generate multidecadal retrospective forecasts, which are long runs where the SST anomalies are prescribed using the same strategy for SST prediction as is used by the real-time forecast system. Such historical runs permit a more realistic assessment of operational model skill and model characteristics relevant to real-time forecasting (Goddard and Mason 2002).

Model skill is determined using temporal anomaly correlation scores and areas beneath the relative operating characteristics curve; (ROC Mason 1982; Stanski et al. 1989; Mason and Graham 2002) for model simulation runs of 30 yr or more. The IRI processes the model output using several approaches to estimate categorical likelihood based on past performance of the particular AGCM, many of which are described in Mason et al. (1999). Additional subjectivity is involved in ascertaining whether the retained model signals are resulting from anomalous SST forcing that seems realistic. For example, if there are known weaknesses in the tropical Indian SST anomaly (SSTA) prediction, added caution will be exercised in forecasting for regions over which tropical Indian SST variability is known to exert influence.

APPENDIX B: SEA SURFACE TEMPERATURE PREDICTION FOR THE TROPICAL OCEANS. The tropical Pacific SSTA has been taken from the NCEP coupled ocean–atmosphere model (Ji et al. 1998). Over the Indian Ocean, SSTA is predicted using a canonical correlation analysis (CCA) model developed at the IRI. The predictors are the recent observations of SSTA in the Indian and tropical Pacific Oceans and also the prediction of SSTA for the tropical Pacific, since much of the Indian Ocean SSTA variability correlates highly with that of the tropical Pacific approximately 3 months earlier (Goddard and Graham 1999). The SSTA of the tropical Atlantic Ocean is obtained from a CCA model developed at the Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) in Brazil (Repelli and Nobre 2003) or is prescribed as damped persistence if the forecast season is one in which the CCA skill is not high for the tropical Atlantic SSTA. The IRI began using the tropical Atlantic SSTA predictions in mid-1998. Potential modifications and refinements to the SSTA prediction tools have been identified, and other operational SST predictions are under consideration.

APPENDIX C: THE RANK PROBABILITY SKILL SCORE. The RPSS measures the cumulative squared error between the categorical forecast probabilities and the observed category relative to some reference forecast (Epstein 1969; Wilks 1995). The RPSS considers the forecast probabilities of all three categories in computing the error with respect to the observation. The computation of RPSS begins with computation of the ranked probability score, or RPS. The RPS is defined as

$$\mathrm{RPS}=\sum_{m=1}^{\mathrm{Ncat}}(\mathrm{CP}_{F_m}-\mathrm{CP}_{O_m})^2,$$

where $N_{\mathrm{cat}} = 3$ for tercile forecasts. The vector CP_{F_m} represents the cumulative probabilities of the forecast up to category m, and CP_{O_m} is the cumulative observed "probability" up to category m. The probability distribution of the observation is 100% for the category that was observed and is zero for the other two categories. The cumulative probability for the observation (CP_O), then, is zero until the observed category is reached, at which time it becomes 1 (see Fig. A1). Low RPS indicates high skill, and vice versa. The RPSS is the RPS of the forecast compared with the RPS of the reference forecast of climatology that assigns 33.3% for each of the tercile categories:

$$\mathrm{RPSS}=1-\frac{\mathrm{RPS}_{fcst}}{\mathrm{RPS}_{ref}},$$

where $\mathrm{RPS}_{\mathrm{fcst}}$ is the RPS for the actual forecast, and $\mathrm{RPS}_{\mathrm{ref}}$ (= $\mathrm{RPS}_{\mathrm{clm}}$) is the RPS of the climatology fore-

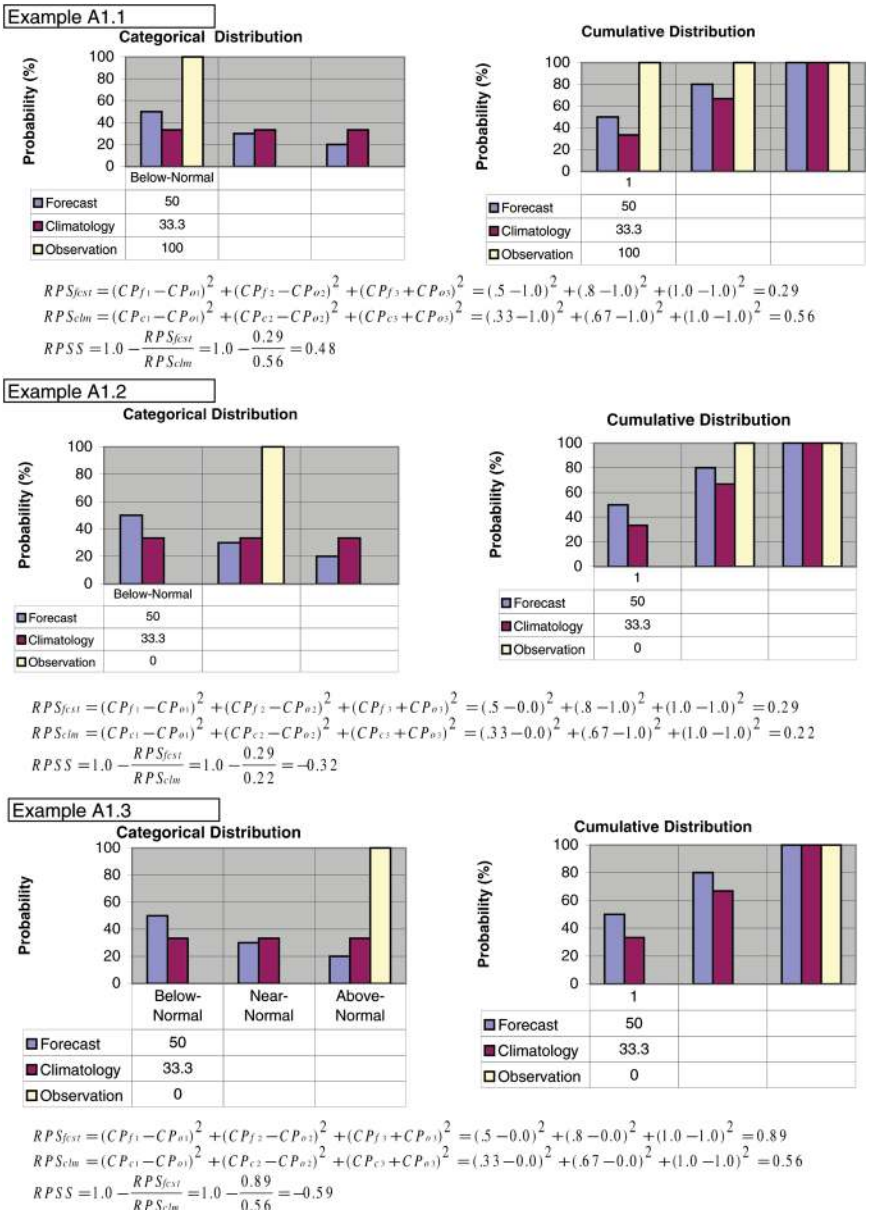

FIG. A1. Histogram plots showing (left) the categorical distributions and (right) the cumulative distributions of the forecast probabilities (blue), the climatology probabilities (red), and the observed "probability" (white). Three different scenarios are shown. The forecast and climatology probabilities are the same, but the observation is different, in each case: 1) observation is below normal; 2) observation is near normal; and, 3) observation is above normal.

cast. The value of RPS_{clm} depends on which category was observed, being lower for the middle category than the two outer categories.

A series of examples of the RPSS are given in Fig. A1. The forecast is the same in each case: 50% probability for below-normal, 30% probability for near-normal, and 20% probability for above-normal precipitation. In the first case (Fig. A1.1), the below-normal category is observed. Thus the dominant forecast category was observed, and the RPSS is positive. Notice that even though the category predicted with a relatively confident probability of 50% was observed, the RPSS is only 48%. In the second case, near-normal conditions were observed. The RPSS value is negative. Note that RPS_{clm} is smaller in this case than it is when the observation falls in one of the outer categories (Figs. A1.1, A1.3). In the last case, above-normal conditions are observed, and the RPSS attains a large negative value (–59%) that is of greater magnitude than the positive value obtained in Fig. A1.1.

REFERENCES

Agrawala, S., K. Broad, and D. G. Guston, 2001: Integrating climate forecasts and societal decision-making: Challenges to an emergent boundary organization. *Science, Technol. Hum. Values,* **26,** 454–477.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak, 2003: Increased automation and use of multimodel ensembling in seasonal climate forecasting at the IRI. *Bull. Amer. Meteor. Soc.,* 1783–1796.

Basher, R., C. Clark, M. Dilley, and M. Harrison, Eds., cited 2001: Coping with the climate: A way forward. A multi-stakeholder review of Regional Climate Outlook Forums concluded at an international workshop. [Available online at http://iri.columbia.edu/outreach/publication/irireport/PretoriaSumRpt2.html].

Beltrando, G., and P. Camberlin, 1993: Interannual variability of rainfall in the eastern Horn of Africa and indicators of atmospheric circulation. *Int. J. Climatol.,* **13,** 533–546.

Bhalme, H. N., S. K. Jadhav, D. A. Mooley, and B. V. Ramana Murthy, 1986: Forecasting of monsoon performance over India. *Int. J. Climatol.,* **6,** 347–354.

Blanford, H. F., 1884: On the connexion of Himalayan snowfall and seasons of drought in India. *Proc. Roy. Soc. London,* **37,** 3–22.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **87,** 367–374.

Derome, J., and Coauthors, 2001: Seasonal predictions based on two dynamical models. *Atmos. Ocean,* **39,** 485–501.

Deutches Klimarechenzentrum, 1992: The ECHAM3 atmospheric general circulation model. DKRZ Tech. Rep. 6, Hamburg, Germany, 184 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Farmer, G., 1988: Seasonal forecasting of the Kenya coast Short Rains, 1901–84. *J. Climatol.,* **8,** 489–497.

Frederiksen, C. S., H. Zhang, R. C. Balgovind, N. Nicholls, W. Drosdowsky, and L. Chambers, 2001: Dynamical seasonal forecasts during the 1997/98 ENSO using persisted SST anomalies. *J. Climate,* **14,** 2675–2695.

Goddard, L., and N. E. Graham, 1999: The importance of the Indian Ocean for simulating rainfall anomalies over eastern and southern Africa. *J. Geophys. Res.,* **104,** 19 099–19 116.

——, and S. J. Mason, 2002: Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Climate Dyn.,* **19,** 619–631, doi:10.1007/s00382-002-0251-y.

——, ——, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.,* **21,** 1111–1152.

Golnaraghi, M., 1997: Applications of seasonal-to-interannual climate forecasts in five U.S. industries: A report to the NOAA's Office of Global Programs. Climate Risk Solutions, Inc., 101 pp. [Available from Climate Risk Solutions, Inc., 25 Thatcher St., Suite #4, Brookline, MA.]

Graham, N. E., 1995: Simulation of recent global temperature trends. *Science,* **267,** 666–671.

Hack, J. J., J. T. Kiehl, and J. W. Hurrell, 1998; The hydrologic and thermodynamic characteristics of the NCEP CCM3. *J. Climate,* **11,** 1179–1206.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.,* **83,** 683–698.

Hastenrath, S., 1995: Recent advances in tropical climate predictions. *J. Climate,* **8,** 1519–1532.

——, M. C. Wu, and P. S. Chu, 1984: Towards the monitoring and predictions of northeast Brazil droughts. *Quart. J. Roy. Meteor. Soc.* **110,** 411–425.

Ji, M., D. W. Behringer, and A. Leetmaa, 1998: An improved coupled model for ENSO predictions and implications for ocean initialization. Part II: The coupled model. *Mon. Wea. Rev.,* **126,** 1022–1034.

Kawamura, R., T. Matsuura, and S. Iizuka, 2001: Role of equatorial asymmetric sea surface temperature anomalies in the Indian Ocean in the Asian summer monsoon and El Niño–Southern Oscillation coupling. *J. Geophys. Rev.,* **106,** 4681–4693.

Kumar, A., W. Wang, M. P. Hoerling, A. Leetmaa, and M. Ji, 2001: The sustained North America warming of 1997 and 1998. *J. Climate,* **14,** 345–353.

——, F. Yang, L. Goddard, and S. Schubert, 2003: Differing trends in the tropical surface temperatures and precipitation over land and oceans. *J. Climate,* in press.

Lall, U., and A. Sharma, 1996: A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.,* **32,** 679–693.

Livezey, R. E., M. Masutani, and M. Ji, 1996: SST-forced seasonal simulation and prediction skill for versions of the NCEP/MRF model. *Bull. Amer. Meteor. Soc.,* **77,** 507–517.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mason, S. J., and L. Goddard, 2001: Probabilistic precipitation anomalies associated with ENSO. *Bull. Amer. Meteor. Soc.,* **82,** 619–638.

——, and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

——, L. Goddard, N. E. Graham, E. Yulaeva, L. Sun, and P. A. Arkin, 1999: The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Amer. Meteor. Soc.,* **80,** 1853–1873.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 19–74.

Mutai, C. C., M. N. Ward, and A. W. Colman, 1998: Towards the prediction of the East Africa short rains based on sea-surface temperature–atmosphere coupling. *Int. J. Climatol.,* **18,** 975–997.

New, M., M. Hulme, and P. D. Jones, 1999: Representing twentieth-century space–time climate variability.

Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate,* **12,** 829–856.

——, ——, and ——, 2000: Representing twentieth-century space–time climate variability. Part II: Development of a 1901–96 monthly grid of terrestrial surface climate. *J. Climate,* **13,** 2217–2238.

NOAA, cited 1999: An experiment in the application of climate forecast: NOAA/OGP activities related to the 1997–1998 El Niño event. NOAA/Office of Global Programs, Washington, DC. [Available online at www.ogp.noaa.gov/enso/retro/ensodoc.htm.]

O'Lenic, E., 1994: A new paradigm for productions and dissemination of the NWS's long lead-time seasonal climate outlooks. *Proc. 19th Annual Climate Diagnostics Workshop,* College Park, MD, NOAA Climate Prediction Center, 408–410.

Oort, A. H., and J. J. Yienger, 1996: Observed interannual variability in the Hadley circulation and its connection to ENSO. *J. Climate,* **9,** 2751–2767.

Peng, P., A. Kumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP–MRF9 and Scripps–MPI ECHAM3 models. *J. Climate,* **13,** 3657–3679.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.,* **130,** 1792–1811.

Repelli, C. A., and P. Nobre, 2003: Statistical prediction of tropical Atlantic sea surface temperature over the tropical Atlantic. *Int. J. Climatol.,* in press.

Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analysis using optimal interpolation. *J. Climate,* **7,** 929–948.

Ropelewski, C. F., J. E. Janowiak, and M. S. Halpert, 1985: The analysis and display of real time surface climate data. *Mon. Wea. Rev.,* **113,** 1101–1106.

Servranckx, R., N. Gagnon, L. Lefaivre, and A. Plante, 1999: Environment Canada seasonal forecasts: Products, methods, and procedures. *Proc. Sixth Workshop on Operational Meteorology,* Halifax, NS, Canada, Environment Canada, 172–176.

Shukla, J., and D. A. Paolino, 1983: The Southern Oscillation and long range forecasting of the summer monsoon rainfall over India. *Mon. Wea. Rev.,* **111,** 1830–1837.

Stanski, H. R., L. J. Wilson, and R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rep. 8, WMO/TD 358, 114 pp.

Walker, G. T., 1923: Correlation in seasonal variations of weather. VIII: A preliminary study of world weather. *Mem. Indian Meteor. Dept.,* **24,** 75–131.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the forth Nordeste of Brazil using eigenvectors of sea-surface temperatures. *Int. J. Climatol.,* **11,** 711–743.

——, ——, 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 Climate Variations,* J. Shukla, Ed., NATO ASI Series, Vol. 16, Springer-Verlag Berlin Heidelberg, 197–216.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 2000: Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–98. *J. Climate,* **13,** 2389–2403.

——, and C. M. Godfrey, 2002: Diagnostic verification of the IRI Net Assessment Forecasts, 1997–2000. *J. Climate,* **15,** 1369–1377.

Xie, P. P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78,** 2539–2558.