Evaluation of IRI's Seasonal Climate Forecasts for the Extreme 15% Tails

ANTHONY G. BARNSTON AND SIMON J. MASON

International Research Institute for Climate and Society, The Earth Institute at Columbia University, Palisades, New York

(Manuscript received 17 August 2010, in final form 28 October 2010)

ABSTRACT

This paper evaluates the quality of real-time seasonal probabilistic forecasts of the extreme 15% tails of the climatological distribution of temperature and precipitation issued by the International Research Institute for Climate and Society (IRI) from 1998 through 2009. IRI's forecasts have been based largely on a two-tiered multimodel dynamical prediction system. Forecasts of the 15% extremes have been consistent with the corresponding probabilistic forecasts for the standard tercile-based categories; however, nonclimatological forecasts for the extremes have been issued sparingly. Results indicate positive skill in terms of resolution and discrimination for the extremes forecasts, particularly in the tropics. Additionally, with the exception of some overconfidence for extreme above-normal precipitation and a strong cool bias for temperature, reliability analyses suggest generally good calibration. Skills for temperature are generally higher than those for precipitation, due both to correct forecasts of increased probabilities of extremely high (above the upper 15th percentile) temperatures associated with warming trends, and to better discrimination of interannual variability. However, above-normal temperature extremes were substantially underforecast, as noted also for the IRI's tercile forecasts.

1. Introduction

The International Research Institute for Climate and Society (IRI) began issuing seasonal forecasts of near-global climate in October 1997, using a two-tiered dynamically based multimodel prediction system (Mason et al. 1999). The standard forecast product, whose quality has been evaluated in depth (Wilks and Godfrey 2002; Goddard et al. 2003; Barnston et al. 2010), contains probabilities of occurrence for the three climatologically equiprobable categories of seasonal total precipitation and mean temperature: below, near, and above normal as defined by the 30-yr base period in use at the time. Probabilistic forecasts for events falling into the lower or upper 15 percentiles of the climatological distribution began being issued in April 1998 and March 2001 for precipitation and temperature, respectively. Based on the same model output as the tercile-based forecasts, they are issued only for the shortest lead time: the 3-month period beginning a half-month following forecast issuance.

Although the lower and upper 15% tails may not necessarily represent near-record mean seasonal conditions (or extreme weather events within the season), probability forecasts for the 15% tails are provided for users particularly sensitive to climate events farther away from the climatic average than can be specifically represented by tercile-based categories.

In the two-tiered dynamical climate prediction methodology (Bengtsson et al. 1993) used for IRI's climate forecasts, a set of SST prediction scenarios is first established, and then a set of atmospheric general circulation models (AGCMs), each consisting of multiple ensemble runs, is forced by the members of the set of predicted SSTs (Mason et al. 1999). During the early 2000s the set of constituent AGCMs expanded, automation increased, and objective multimodel ensembling methodologies were implemented (Rajagopalan et al. 2002; Barnston et al. 2003; Robertson et al. 2004). Following production of the purely objective forecast probabilities for the standard tercile-based forecasts, final minor subjective modification is carried out by the forecasters (Barnston et al. 2010), leading to more probabilistically reliable forecasts.

Although forecasts for the 15% extremes are based on the same model output as the tercile-based forecasts, they are issued in a less quantitative format. While the

Corresponding author address: Anthony Barnston, International Research Institute for Climate and Society, P.O. Box 1000, 61 Rt. 9W, Columbia University, Palisades, NY 10964-8000.
E-mail: tonyb@iri.columbia.edu

DOI: 10.1175/WAF-D-10-05009.1

tercile forecasts are expressed in increments of 5%, just three gradations of probability enhancement above the 15% climatological level are defined for the 15% extremes forecasts: *slightly enhanced* (defined by probabilities of 25%–40%), *enhanced* (40%–50%), and *greatly enhanced* (≥50%). Together with the climatological *neutral default* (<25%), this forecast format was elected both to make the forecasts more easily understood by users, and because of the greater uncertainty associated with forecast probabilities in the outer portions of the climatological distribution. No decreased probabilities of extreme conditions are explicitly forecast in either tail. In developing forecasts for the 15% tails, the forecasters use a combination of guidance consisting of the postprocessed model output and its multimodel combination, extrapolation from the tercile-based forecasts (which are usually developed first)[1], and subjective judgment. In this paper we evaluate only the final issued forecasts, to which the user community has access, and not the guidance tools used to formulate them. An example of a forecast map for extreme precipitation is shown in Fig. 1, along with its corresponding standard tercile-based probability forecast.

In section 2 the verification data and procedures are defined, in section 3 the verification results are presented, and a summary and some concluding remarks are provided in section 4.

2. Data and methods

a. Verification data

For temperature verification, the 2° gridded global Climate Anomaly Monitoring System (CAMS) dataset from the National Oceanic and Atmospheric Administration (NOAA) (Ropelewski et al. 1985) is used. For precipitation, 2.5° gridded data from the Climate Research Unit (CRU) of the University of East Anglia for 1961–78 (New et al. 2000; Mitchell and Jones 2005) are used, and the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP; Xie and Arkin 1997) dataset is used from 1979 through 2009.[2] Consistency tests between the two datasets during the overlapping period indicate minor biases in the mean, and somewhat larger biases in variance, with the CRU data having lower variance. The variance bias slightly affects the 15th and 85th percentiles when the 1961–90 climatology base period was used through mid-2001, but has negligible effects when later base periods (1969–98 from mid-2001 through 2002, and 1971–2000 since January 2003) were used. Accounting for a change from a quarterly schedule of forecast issuance before mid-2001 to a monthly schedule thereafter, seasonal extremes forecasts were issued for precipitation for 113 target periods beginning April–June 1998, and for temperature for 101 target periods beginning April–June 2001. The ending target season for both variables is December–February 2009/10.

b. Methods

For the purposes of assessments of reliability, the forecast probability at each grid square is regarded as a value indicative of the rank of its level of enhancement: 1 for no enhancement (i.e., the climatological probability of <25%), and 2, 3, or 4 for each of the respective progressively increasing enhancement levels (25%–40%, 40%–50%, and >50%). This ordinal representation is used because the probability ranges of the categories are too wide to perform a more rigorously quantitative diagnosis. Therefore, we do not use verification measures intended for more precisely defined probability forecasts (e.g., measures related to the ignorance score, Brier score, or the quantitative outputs of a reliability diagnosis). Our goal is to assess the degree to which the forecasts successfully indicate increases in the frequency of occurrence of extremes. Thus, we examine reliability within an ordinal context, using the actual probability ranges only to check for obvious inconsistencies with the observed frequencies of occurrence, and to develop a quasi-quantitative reliability plot. To assess discrimination in the forecasts more specifically, we compute relative operating characteristic (ROC) areas (Mason 1982), which require only ordinally defined probability forecasts. Similarly, the ordinal probability bins allow for the calculation of resolution scores (in fact, there is no implied ordering of the probability bins in the resolution score).

When one of the 15% extreme tails is forecast with an enhanced probability, the probabilities for the opposite extreme and for the larger middle category are not explicitly given in the issued forecasts. Therefore, the reliability assessment given below pertains only to forecast probabilities assigned directly to one of the 15% extremes. However, for ROC evaluation the results are so heavily dominated by the predominance of climatological forecasts that an assumption is made that progressive increases in the probability of one extreme imply progressive decreases in the probability of the opposite extreme. This assumption is compatible with the ordinal framework used in ROC, where probability values themselves are not

[1] Specifically, a Gaussian fit is made to the tercile probabilities, and this probability density function is used to determine the probabilities for the 15% tails.

[2] For the five final months beginning in October 2009, CMAP data were unavailable, and Climate Anomaly Monitoring System–OLR Precipitation Index (CAMS–OPI) rainfall data (Janowiak and Xie 1999) were used instead.

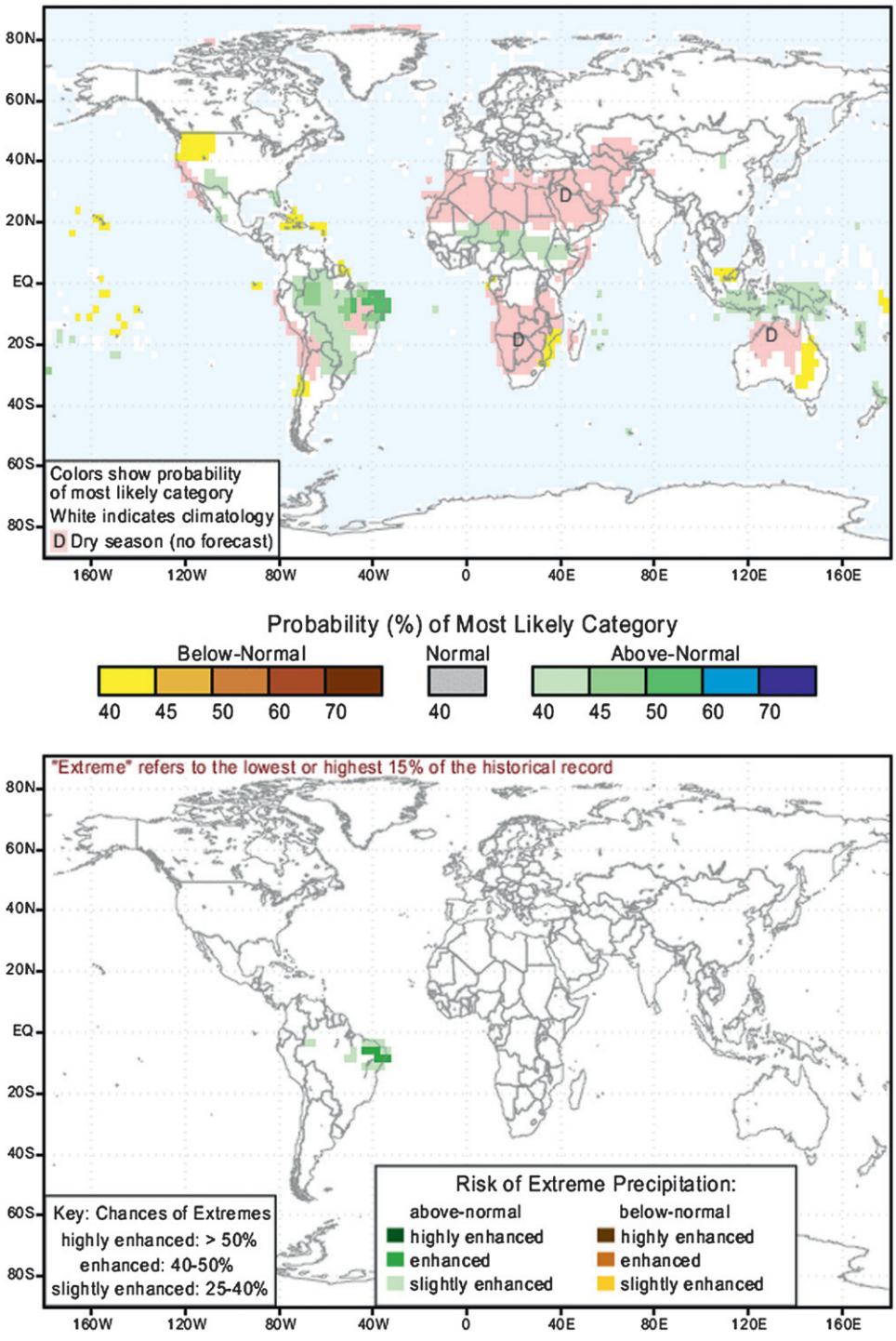

FIG. 1. (bottom) Example of a forecast for the 15% precipitation extremes issued in May 2009 for June–August 2009. An area of enhanced (40%–50%) probability for above-normal rainfall is indicated in part of northeast Brazil, surrounded by some area of slightly enhanced (25%–40%) probability. (top) The corresponding standard tercile-based probability forecast.

TABLE 1. Average of percentage areal coverage of forecasts of nonclimatological probabilities for the 15% extremes for precipitation and temperature over the globe and in the tropics (25°N–25°S). For the total ($p \geq 0.25$), the portions of coverage in each extreme are shown.

Variable	Domain (land only)	Slightly enhanced ($0.25 \leq p < 0.40$)	Enhanced ($0.40 \leq p < .50$)	Greatly enhanced ($p \geq 0.50$)	Total ($p \geq 0.25$) (lower, upper)
Precipitation	Globe	0.86	0.13	0.04	1.03 (0.48, 0.55)
	Tropics	1.73	0.25	0.07	2.05 (0.94, 1.11)
Temperature	Globe	2.98	0.70	<0.005	3.68 (0.11, 3.57)
	Tropics	5.03	1.00	<0.005	6.03 (0.21, 5.82)

used, but rather just their relative rank ordering. This assumption also implies that forecast distributions having highly non-Gaussian shapes (e.g., bimodal) or greatly varying spreads, which could violate the opposing directions of probability change on opposite tails, are rare. Since such markedly non-Gaussian forecasts have not been issued in the tercile-based IRI forecasts, the assumption is reasonable.

Reliability, or attributes, diagrams (Murphy 1973; Wilks 2006) show the correspondence of the full range of issued forecast probabilities and their associated relative frequencies of observed occurrence. Ideally, the forecast probabilities would closely match the observed relative frequencies of occurrence for each of the lower 15% and upper 15% climatological categories. The diagrams can reveal forecast characteristics such as probabilistic bias, forecast over- (under-) confidence, and forecast sharpness. Here, observed relative frequencies are examined for the four ordinal forecast probability bins, providing a rough indication of forecast reliability and resolution.

The ROC area is the area under the curve of hit rate versus false-alarm rate on a ROC plot (Mason 1982). The ROC plot shows the cumulative hit rate against cumulative false-alarm rate for progressively decreasing forecast probabilities for an event to occur (e.g., the event of exceeding the upper 15th percentile for precipitation). A favorable ROC plot would show a higher hit rate than false-alarm rate for cases having the highest probabilities for the event, but with increasingly less frequent hits and more frequent false alarms as forecasts with lower probabilities are added into the cumulative tally. With a possible range of 0%–100%, a 50% rate of correct discrimination is expected by chance and reflects 0 forecast skill (Mason and Weigel 2009). ROC measures discrimination alone, without penalty for poor probability calibration. The use of ordinal forecast probability bins here does not affect the computation of the ROC area, as the results would be identical regardless of what probabilities are assigned to the bins, provided that their rank ordering matches that of the ordered bins.

To address significance for the ROC areas, Monte Carlo simulations are performed in which the years of the observations are randomly permuted among the 12 (9) yr for precipitation (temperature), while the ordering of the months within each year remains intact to preserve the integral time scale of the forecast and observed data (i.e., to maintain the temporally correlated climate responses within an ENSO cycle). Five thousand randomizations are conducted. Similarly, the sampling errors in the ROC areas are represented using confidence intervals, determined using a bootstrapping technique. In the bootstrapping 1-yr segments of individual forecasts are randomly resampled with replacement, while the true forecast–observation pairs remain intact (Wilks 2006; Mason 2008). A sample size of 5000 is used.

3. Results

Reliability analyses and ROC scores are calculated as averages of the results including all relevant grid squares, where each square is area weighted by the cosine of its latitude.

a. Coverage of nonclimatological probabilities

Issuance of enhanced probabilities for the 15% extremes has been conservative (Table 1). For precipitation, the areal coverage of nonclimatological probabilities has averaged approximately 1.0% of the global land area and 2.0% of the tropical land area (25°N–25°S), while for temperature the coverage areas have been approximately 3.7% and 6.0%, respectively. Climatologically one would expect 30% of the globe to experience an extreme of one sign or the other. Additionally, more than 80% of the forecasts for enhanced probabilities have been for the weakest level of enhancement (25%–40%) for both precipitation and temperature, for both global and tropical domains (Table 1). Figure 2 shows the geographical distribution of the percentage frequency of issuance of nonclimatological probabilities for precipitation and temperature. These results indicate the most frequent issuance of enhanced probabilities for extreme precipitation in Indonesia, the Philippines, tropical Pacific islands along the immediate equator, in far western Africa, and near the coast of northeast South America. Precipitation

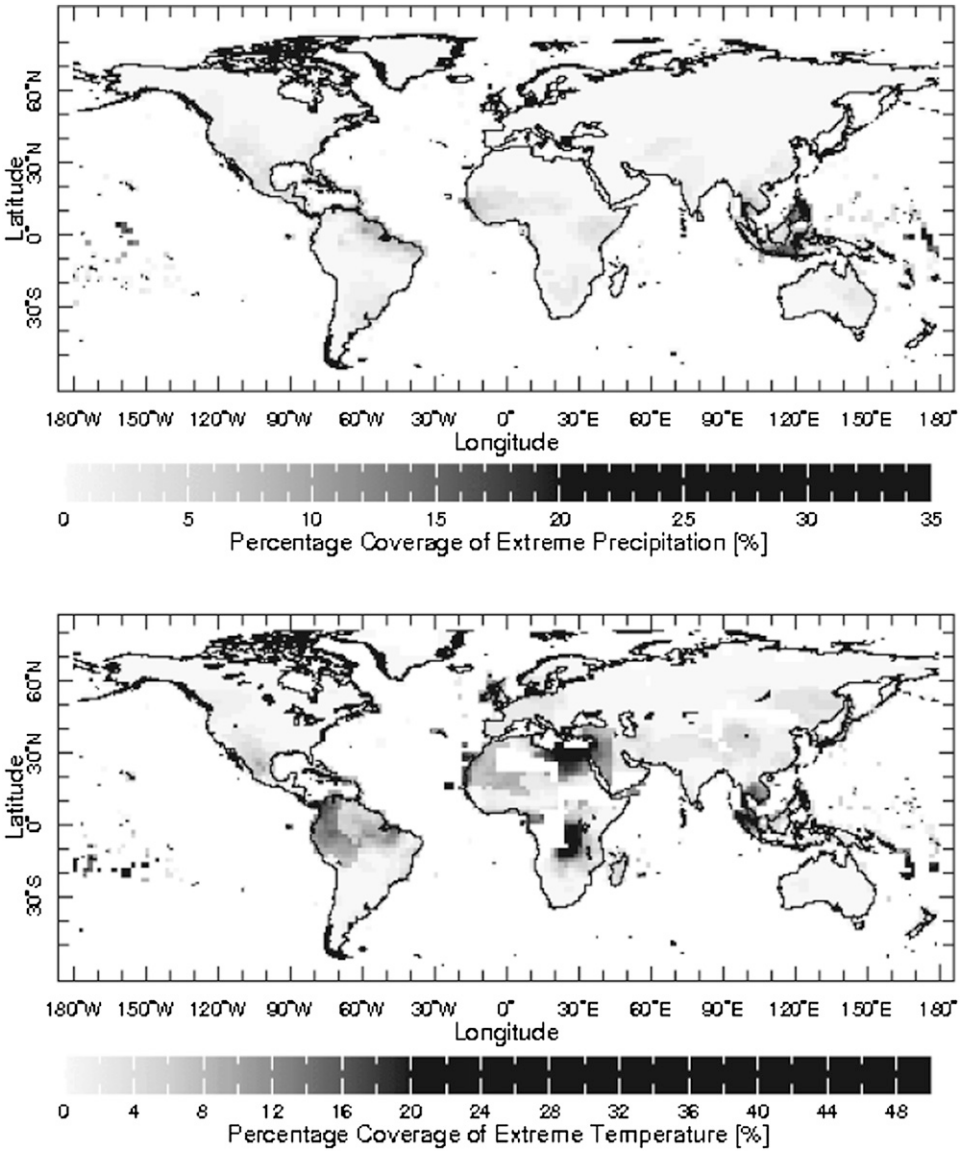

FIG. 2. Geographical distribution of the percentage frequency of issuance of forecasts for nonclimatological probabilities for the 15% extremes for (top) precipitation and (bottom) temperature. Completely white areas represent oceans and other larger water bodies and for temperature, land regions having substantial proportions of missing observational data.

extremes were infrequently forecast in the middle and high latitudes. Temperature extremes were forecast most frequently in the region surrounding the eastern Mediterranean Sea (Egypt, northern Saudi Arabia, western Middle East), central southern Africa, western Africa, northern South America, southeast Asia and western Indonesia, and many of the South Pacific islands. The coverage for both precipitation and temperature varies seasonally, being somewhat greater from late northern autumn through late northern spring, and less during northern summer and early autumn (not shown). This seasonal cycle of coverage is likely related to the seasonal distribution of confidence in the climate effects associated with ENSO episodes, and coincides with the globally averaged skill of IRI's standard tercile-based forecasts (Barnston et al. 2010).

b. Reliability and resolution

Reliability plots for forecasts in the tropics over the 11-yr forecast period, aggregated over all seasons and area-weighted grid points, are shown in Fig. 3 for precipitation and temperature for the lower and upper 15%

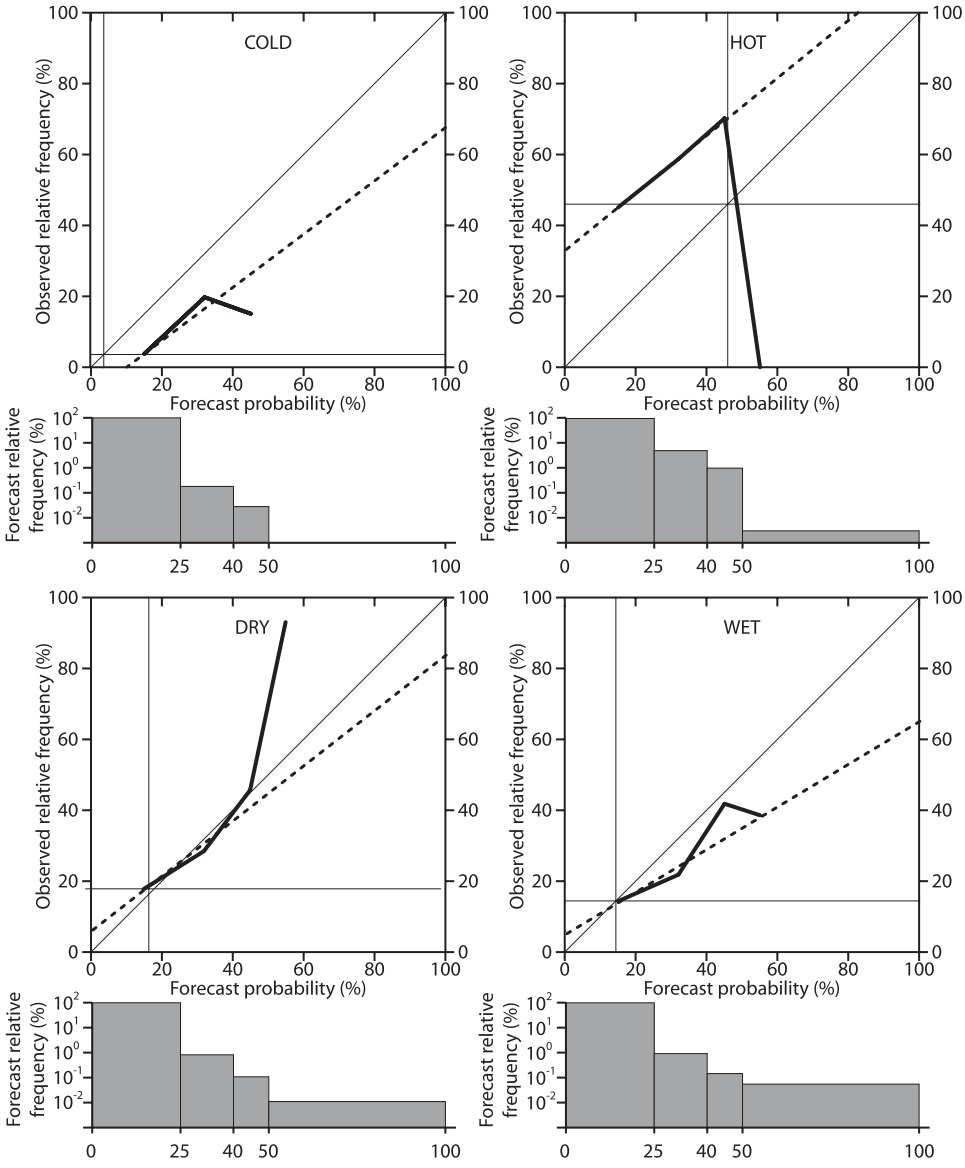

FIG. 3. Reliability plots for forecasts of (top) temperature and (bottom) precipitation extremes in the tropics (25°N–25°S). The straight 45° line represents ideal reliability. The dashed line is the least squares linear regression fit to the points forming the reliability curve, weighted by the sample sizes represented by each point. Horizontal and vertical lines are drawn at the observed relative frequencies for the study period. Forecast probabilities are plotted at the midpoints of their respective probability intervals, except ''neutral default'' is plotted at 15% and ''extremely enhanced'' at 55% because values greater than 60% were never indicated. Subpanels below each chart show the percentage frequencies with which the four forecast probability categories were forecast on a logarithmic ordinate scale.

categories individually. Observed relative frequencies associated with each forecast category are shown in the main panels of Fig. 3 for the tropics, and are also indicated in Table 2 for both the tropics and the globe. A regression line is drawn in Fig. 3 to summarize the reliability curve when weighted by the number of forecast cases. The panels in Fig. 3 beneath the reliability plots indicate the relative frequencies of issuance among the four forecast probability categories. As suggested earlier, Fig. 3 is intended as a quasi-quantitative plot, because the probability intervals are deemed too wide, and there are too few of them, to justify an accurate assignment of their single representative probability values. Our evaluation targets the overall features of the forecasts—ones that can be seen clearly with only rough precision.

TABLE 2. Diagnostics associated with reliability plots and ROC areas, for each of the 15% extreme tails for precipitation and temperature. Observed (Obs) relative (Rel) frequencies (Freq) are given for all forecast cases as well as for each of the four forecast (Fcst) categories. ROC area includes contributions from progressively lower probabilities in the tail opposite that being forecast with progressively higher probabilities (see text). Superscripts to the ROC areas indicate Monte Carlo–based significance levels, as percentages (e.g., 0.1 indicates a significance level of ≤0.001). The 95% confidence interval (CI) for the ROC area is indicated (see text).

Variable	Domain	Extreme	Coverage (%)	Obs Rel Freq (%) Uncon 0–100	Obs Rel Freq (%) Fcst 1 <25	Obs Rel Freq (%) Fcst 2 25–40	Obs Rel Freq (%) Fcst 3 45–50	Obs Rel Freq (%) Fcst 4 >50	ROC area (95% CI)
Precipitation	Globe	Wet	0.55	15.1	15.0	21.5	40.6	40.5	$0.504^{0.1}$ (0.502–0.508)
		Dry	0.48	19.0	18.5	28.5	44.1	92.9	$0.504^{0.1}$ (0.502–0.508)
	Tropics	Wet	1.11	14.4	14.3	21.8	41.7	38.5	$0.509^{0.1}$ (0.505–0.515)
		Dry	0.94	17.8	17.7	28.4	45.6	92.9	$0.508^{0.1}$ (0.504–0.515)
Temperature	Globe	Hot	3.57	39.5	38.9	56.0	65.1	0.0	$0.514^{0.1}$ (0.508–0.523)
		Cold	0.11	5.1	4.9	25.4	27.3		$0.516^{0.1}$ (0.511–0.523)
	Tropics	Hot	5.82	45.9	44.9	58.6	70.1	0.0	$0.519^{0.1}$ (0.510–0.532)
		Cold	0.21	3.8	3.8	19.7	15.1		$0.524^{0.1}$ (0.516–0.532)

While the precipitation resolution and reliability appear to be generally satisfactory, some overconfidence is noted, particularly for extreme above-normal precipitation where the observed relative frequency is only slightly over 20% when the slightly enhanced (25%–40%) category is forecast; this is only about 7% higher in frequency than when the neutral default (<25%) is forecast.[3] Correspondingly, the slope of the least squares linear regression fit to the four points in the bottom-right panel of Fig. 3 is somewhat shallower than the ideal 45° line. The frequency of issuance of the given forecast probability categories for the 15% tails (Tables 1 and 2, and lower subpanels in Fig. 3) for precipitation shows an overwhelming majority of climatological probability forecasts for the tropics (98.0% of cases), and to an even greater extent for the globe (99.0%). The resolution for the precipitation forecasts is 0.0004 and 0.0003 for the dry and wet extreme categories, respectively, in the tropics, and roughly one-half of these values for the globe. These values are very small because of the overwhelming preponderance of climatological forecasts.

For temperature, slight overconfidence is also present, but to a lesser extent than for precipitation, with changes in observed relative frequency near 20% for the globe and 15% for the tropics between cases when climatology was forecast and when the first level of probability enhancement (25%–40%) was forecast. Resolution for the temperature forecasts is 0.0001 and 0.0016 for the cold and hot extreme categories, respectively, in the tropics, and 0.0002 and 0.0012, respectively, for the globe. Thus, higher resolution is indicated for temperature than for precipitation, but primarily for the hot extreme category, which was forecast with enhanced probability most frequently.

However, as indicated in Table 2, above-normal extreme temperatures were markedly underforecast: globally (within the tropics), the hot 15% extreme was observed on about 40% (46%) of occasions, but the forecasts implied frequencies of only about 16% (16%) (i.e., only marginally more than climatology) because of the preponderance of climatological forecasts. Hence, the regression-fitted reliability curves for temperature (Fig. 3) show substantial over- (under-) forecasting of cold (hot) extremes. The results for the standard tercile-based forecasts indicated the same problem (Barnston et al. 2010). However, the extremes forecasts did at least imply a much larger frequency of hot compared to cold extremes: hot extremes globally (in the tropics) were observed about 8 (12) times as frequently as cold extremes, while the forecasts implied a value of about 32 (28). The great imbalance in the observations of upper versus lower extreme 15% categories reflects the magnitude of the low-frequency variability, including specifically a global warming signal.

c. *ROC area*

The last column in Table 2 shows ROC area results for the globe and the tropics, indicative of probabilistic discrimination skill. Because enhanced probabilities for the 15% extremes were forecast sparingly, a very large proportion of the forecasts are indistinguishable (even after the default climatological probability is subdivided by assuming forecasts of diminished probability for the opposite extreme), and the ROC areas are therefore damped severely toward 0.50, implying little ability to

[3] The change in observed relative frequency from the climatology forecast category to the first forecast enhancement level is most critical in assessing forecast confidence, because the forecast categories of greater enhancement level were forecast much less frequently for both precipitation and temperature (Table 1).

TABLE 3. ROC areas by region. The 95% confidence interval for the ROC area, based on a bootstrap test, is shown in parentheses. The *p* value for the ROC area being 0.5 or lower is shown and is based on an independent Monte Carlo test. ROC areas with *p* values of 0.05 or better are shown in boldface. Occasional minor inconsistencies between the Monte Carlo significance tests and bootstrap confidence intervals reflect sampling errors and violations in the block sampling procedures used (which are more noticeable for temperature than for precipitation).

	Precipitation			
	Dry (lower 15%)		Wet (upper 15%)	
	ROC (95% CI)	*p* value	ROC (95% CI)	*p* value
Southeast Asia	0.501 (0.493–0.512)	0.364	0.501 (0.482–0.516)	0.464
Indonesia and vicinity	**0.514** (0.500–0.531)	0.004	**0.521** (0.508–0.540)	<0.001
Philippines	**0.521** (0.510–0.537)	0.006	**0.540** (0.516–0.568)	0.006
Western tropical Pacific Islands	**0.547** (0.515–0.582)	<0.001	**0.535** (0.509–0.572)	<0.001
Eastern tropical Pacific Islands	**0.519** (0.513–0.525)	<0.001	**0.519** (0.507–0.538)	<0.001
Africa	**0.503** (0.501–0.506)	0.037	**0.503** (0.501–0.507)	0.026
Southern United States–Caribbean–Mexico–Central America	**0.510** (0.501–0.524)	0.002	**0.506** (0.500–0.514)	0.020
Northern South America	**0.513** (0.503–0.527)	0.043	**0.511** (0.502–0.525)	0.008
Tropics	**0.508** (0.504–0.515)	<0.001	**0.509** (0.505–0.515)	<0.001
Globe	**0.504** (0.502–0.508)	<0.001	**0.504** (0.502–0.508)	<0.001
	Temperature			
	Cold (lower 15%)		Hot (upper 15%)	
	ROC (95% CI)	*p* value	ROC (95% CI)	*p* value
Southeast Asia	**0.518** (0.504–0.539)	0.037	0.502 (0.483–0.529)	0.426
Indonesia and vicinity	**0.521** (0.511–0.534)	<0.001	**0.527** (0.512–0.546)	<0.001
Philippines	**0.529** (0.513–0.547)	0.002	**0.533** (0.511–0.560)	0.002
Western tropical Pacific Islands	0.518 (0.415–0.562)	0.246	0.508 (0.490–0.533)	0.244
Eastern tropical Pacific Islands	0.539 (0.435–0.567)	0.087	**0.575** (0.533–0.626)	<0.001
Africa	**0.533** (0.511–0.555)	<0.001	**0.510** (0.498–0.527)	0.037
Southern United States–Caribbean–Mexico–Central America	**0.520** (0.508–0.534)	<0.001	**0.518** (0.507–0.533)	0.023
Northern South America	0.542 (0.513–0.570)	0.170	0.523 (0.496–0.553)	0.094
Tropics	**0.524** (0.516–0.532)	<0.001	**0.519** (0.510–0.532)	<0.001
Globe	**0.516** (0.511–0.523)	<0.001	**0.514** (0.508–0.523)	<0.001

discriminate extremes from nonextremes. ROC areas are between 0.50 and 0.51 for precipitation, and between 0.51 and 0.52 for temperature. Results for temperature are better partly because nonclimatological probabilities were issued roughly 3 times as frequently as they were for precipitation, permitting more opportunity for contributions toward a ROC area >0.5.

The smallness of the exceedance of the ROC areas over 0.5 raises the issue of their statistical significance. Monte Carlo results (shown by superscripts to the ROC areas in Table 2) indicate statistical significance levels of <0.001 for both extremes and for both variables for the global and tropical domains, providing evidence that the chances of these skill levels emerging by accident are remote. Similarly, the bootstrap confidence intervals indicate minimal sampling uncertainty in the ROC areas, and the fact that none of the intervals straddles 0.5 confirms the skill of the forecasts.

Because the geographical coverage of grid squares receiving a meaningful sample of forecasts for enhanced probabilities of the 15% extremes is limited, a map showing the geographical variation of ROC skill would have large blank areas and areas with excessively noisy results. Therefore, we examine the geographical variability of ROC skill for discrete regions of adequate size and/or coverage as indicated in Fig. 2. The resulting regionally aggregated ROC areas are shown in Table 3, along with their 95% confidence intervals based on the bootstrap resampling, and their statistical significance based on the Monte Carlo method. For precipitation, relatively high and statistically significant skill is found in the Philippines and in the western tropical Pacific islands, particularly for wet forecasts for the former and for dry forecasts for the latter. Statistically significant but smaller ROC areas appear in the eastern tropical Pacific islands, and in Indonesia for wet forecasts. With the exception of Southeast Asia, ROC areas for all of the selected regions are statistically significant at the 5% level or better, despite that some are only slightly greater than 0.5, such as Africa and the globe as a whole. For temperature, relatively large ROC areas for the extreme tails are found in the eastern tropical Pacific islands, but only forecasts for the hot extreme achieve statistical significance. Skillful forecasts are seen also in Africa for cold

extremes and to a somewhat lesser extent in Indonesia. As might be expected, regions having relatively high frequencies of forecasts of enhanced probabilities (Fig. 2) tend to have higher ROC skill levels.

4. Summary and conclusions

The IRI has issued probabilistic forecasts of near-global seasonal mean temperature and total precipitation extremes (defined by the lower and upper 15% tails) since early 1998 for precipitation and since mid-2001 for temperature. The forecasts have been based primarily on a two-tiered, dynamically based prediction system where a set of SST prediction scenarios is made, serving as prescribed lower boundary conditions for integrations of ensembles from a set of AGCMs. Seven AGCMs have been used since 2004, producing well over 100 forecast ensembles that are postprocessed and merged into final probability forecasts.

Forecasts for the 15% extremes were issued conservatively, resulting typically in small spatial coverage, with forecasts issued mostly in the tropics and subtropics, and with a preponderance of the weakest category of enhanced probability (i.e., 25%–40%). This cautiousness may have been exacerbated by the coarseness of the issued probability bins, making forecasters reluctant to forecast even the weakest level of enhanced probability, whose minimum probability (25%) is two-thirds greater than the climatological probability. Indeed, using a Gaussian fit to the corresponding tercile forecast probabilities, a probability of approximately 50%–55% is required for an extreme tercile category to achieve a probability of 25% for the corresponding 15% tail. The resulting preponderance of forecasts of the climatological category (lower subpanels in Fig. 3), and the consequently low forecast sharpness, strongly affects measures of the forecasts' quality. Nonetheless, within the set of extremes forecasts that *were* issued, results indicate largely satisfactory resolution and favorable calibration, with two notable exceptions: 1) forecasts for extreme above-normal precipitation were somewhat overconfident and 2) above-normal temperature extremes were substantially underforecast.

Skill levels for temperature average somewhat higher than those for precipitation, due to the correct recognition of increasing warmth within the approximately 1-decade period, and better discrimination of interannual variability within the period. Precipitation skill, based more exclusively on correctly discriminated interannual variability, may be somewhat hindered by spatially noisier patterns than those for temperature (Gong et al. 2003) under comparably predictable associated large-scale circulation anomalies.

Although temperature extremes were forecast with greater coverage and skill, the warm extreme was substantially underforecast, as noted also for IRI's tercile-based forecasts (Wilks and Godfrey 2002; Barnston et al. 2010). This bias had been seen also in the climate forecasts made by NOAA's Climate Prediction Center (Wilks 2000; Livezey and Timofeyeva 2008), despite the fact that recent trend indicators are used in developing their forecasts (Huang et al. 1996; O'Lenic et al. 2008). Underforecasting warmth is likely a result, at least in part, of the dynamical prediction systems used at both IRI and CPC that use fixed greenhouse gas concentrations (at late 1980s levels) rather than time-varying concentrations that keep pace with observed increases (Doblas-Reyes et al. 2006; Liniger et al. 2007). Models at many institutions, including IRI, are beginning to use flexible concentrations to help remedy this problem. In the IRI forecast system, the failure of both the SST and atmospheric models to reproduce the full strength of the global warming signal results in the loss of skill not only because current temperature forecasts are biased toward cold, but also because the weakened skill over the training period for the recalibration and combination schemes tends to dampen the models' signals, resulting in smaller probability shifts.

Improvements are being implemented in IRI's forecast system. First, its newly configured forecasts issue the full probability distribution, making possible probabilities for flexibly defined categories. The new system performs multivariate rather than merely local calibrations of individual model forecast outputs prior to multimodel combination (e.g., Landman and Goddard 2002; Tippett et al. 2003; Ndiaye et al. 2009). Additionally, single-tiered (coupled) models are being introduced, and incorporation of time-varying greenhouse gas settings is under way. These changes are expected to increase forecast quality in a broad sense, and have already resulted in larger coverage areas over which enhanced probabilities for the upper or lower 15% tails are being issued.

Acknowledgments. This work was funded by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA050AR4311004 and NA10OAR4310210). The views expressed are those of the authors and do not necessarily reflect the views of NOAA or its subagencies.

REFERENCES

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.,* **84,** 1783–1796.

——, S. Li, S. J. Mason, D. G. DeWitt, L. Goddard, and X. Gong, 2010: Verification of the first 11 years of IRI's seasonal climate forecasts. *J. Appl. Meteor. Climatol.,* **49,** 493–520.

Bengtsson, L., U. Schlese, E. Roeckner, M. Latif, T. P. Barnett, and N. E. Graham, 1993: A two-tiered approach to long-range climate forecasting. *Science,* **261,** 1027–1029.

Doblas-Reyes, F. J., R. Hagedorn, T. N. Palmer, and J. J. Morcrette, 2006: Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophys. Res. Lett.,* **33,** L07708, doi:10.1029/2005GL025061.

Goddard, L., A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's ''Net Assessment'' seasonal climate forecasts: 1997–2001. *Bull. Amer. Meteor. Soc.,* **84,** 1761–1781.

Gong, X., A. G. Barnston, and M. N. Ward, 2003: The effect of spatial aggregation on the skill of seasonal precipitation forecasts. *J. Climate,* **16,** 3059–3071.

Huang, J., H. M. Van den Dool, and A. G. Barnston, 1996: Long-lead seasonal temperature prediction using optimal climate normals. *J. Climate,* **9,** 809–817.

Janowiak, J. E., and P. Xie, 1999: CAMS–OPI: A global satellite–rain gauge merged product for real-time precipitation monitoring applications. *J. Climate,* **12,** 3335–3342.

Landman, W. A., and L. Goddard, 2002: Statistical recalibration of GCM forecasts over southern Africa using model output statistics. *J. Climate,* **15,** 2038–2055.

Liniger, M. A., H. Mathis, C. Appenzeller, and F. J. Doblas-Reyes, 2007: Realistic greenhouse gas forcing and seasonal forecasts. *Geophys. Res. Lett.,* **34,** L04705, doi:10.1029/2006GL028335.

Livezey, R. E., and M. M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts—Insights from a skill analysis. *Bull. Amer. Meteor. Soc.,* **89,** 843–854.

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

Mason, S. J., 2008: Understanding forecast verification statistics. *Meteor. Appl.,* **15,** 31–40.

——, and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.,* **137,** 331–349.

——, 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.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.,* **25,** 693–712.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Ndiaye, O., L. Goddard, and M. N. Ward, 2009: Using regional wind fields to improve general circulation model forecasts of July–September Sahel rainfall. *Int. J. Climatol.,* **29,** 1262–1275, doi:10.1002/joc.1767.

New, M., M. Hulme, and P. D. Jones, 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.

O'Lenic, E. A., D. A. Unger, M. S. Halpert, and K. S. Pelman, 2008: Developments in operational long-range climate prediction at CPC. *Wea. Forecasting,* **23,** 496–515.

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.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.,* **132,** 2732–2744.

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.

Tippett, M. K., M. Barlow, and B. Lyon, 2003: Statistical correction of central southwest Asia winter precipitation simulations. *Int. J. Climatol.,* **23,** 1421–1433.

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

——, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 648 pp.

——, 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 estimations, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78,** 2539–2558.