Verification of the First 11 Years of IRI's Seasonal Climate Forecasts

ANTHONY G. BARNSTON, SHUHUA LI, SIMON J. MASON, DAVID G. DEWITT, LISA GODDARD, AND XIAOFENG GONG*

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

(Manuscript received 15 June 2009, in final form 23 September 2009)

ABSTRACT

This paper examines the quality of seasonal probabilistic forecasts of near-global temperature and precipitation issued by the International Research Institute for Climate and Society (IRI) from late 1997 through 2008, using mainly a two-tiered multimodel dynamical prediction system. Skill levels, while modest when globally averaged, depend markedly on season and location and average higher in the tropics than extratropics. To first order, seasons and regions of useful skill correspond to known direct effects as well as remote teleconnections from anomalies of tropical sea surface temperature in the Pacific Ocean (e.g., ENSO related) and in other tropical basins. This result is consistent with previous skill assessments by IRI and others and suggests skill levels beneficial to informed clients making climate risk management decisions for specific applications. Skill levels for temperature are generally higher, and less seasonally and regionally dependent, than those for precipitation, partly because of correct forecasts of enhanced probabilities for above-normal temperatures associated with warming trends. However, underforecasting of above-normal temperatures suggests that the dynamical forecast system could be improved through inclusion of time-varying greenhouse gas concentrations. Skills of the objective multimodel probability forecasts, used as the primary basis for the final forecaster-modified issued forecasts, are comparable to those of the final forecasts, but their probabilistic reliability is somewhat weaker. Automated recalibration of the multimodel output should permit improvements to their reliability, allowing them to be issued as is. IRI is currently developing single-tier prediction components.

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 forecasts are probabilistic with respect to the occurrence of 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. The forecasts were issued quarterly for the two upcoming consecutive 3-month periods from October 1997 until June 2001, after which time they were issued monthly for the same two lead times, but additionally for the two intermediate overlapping 3-month periods. For most of the history of the IRI's forecasts, they have been issued approximately one-half month prior to the beginning of the first 3-month forecast period.[1] We define the lead time as the time between issuance and the start of the targeted period; since June of 2001, forecasts were issued at 0.5-, 1.5-, 2.5-, and 3.5-month lead times.

An evaluation of the performance of IRI's seasonal forecasts from 1997 through 2001 was presented in Goddard et al. (2003). Forecast skills were found to be positive for the seasons and regions known to have intrinsic predictability, aided in particular by the strong ENSO events from mid-1997 through mid-2000. In this paper forecast skills during the longer period of late 1997 through 2008 are described.

In section 2 the methodology used to produce IRI's precipitation and temperature forecasts, and the forecast format, are outlined. In section 3 the verification data

* Current affiliation: Swiss Re Financial Services Corporation, New York, New York.

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

DOI: 10.1175/2009JAMC2325.1

[1] Prior to 2000, forecasts were issued during the first week of the first predicted period.

and procedures are defined, and in section 4 skills of the IRI's forecasts are examined by region and season and in the context of the ENSO state and a strong multidecadal warming trend. Skills of the issued forecasts are compared with those of the objective guidance of the numerical prediction tools. Section 5 provides a summary and suggests possible improvements for IRI's climate forecasts.

2. Climate prediction methodology

The IRI's prediction methodology has been primarily dynamical, using a two-tiered system (Bengtsson et al. 1993) in which a set of SST predictions is first established, and then a set of atmospheric general circulation models (AGCMs), each consisting of multiple ensemble runs, is forced by the set of predicted SSTs (Mason et al. 1999). The use of multiple SST scenarios accounts for uncertainty in the SST predictions, yielding more realistic levels of uncertainty in the temperature and precipitation forecasts than would be produced from a single (but imperfect) SST scenario.

a. SST prediction

The precise details of the method of deriving the SST predictions have evolved during the 11 years of IRI's forecasts, but use of both persisted SST anomalies and one or more scenarios of evolving SST predictions based on a combination of dynamical and statistical models have been consistent features. In all scenarios, the SST forecasts in the extratropics (outside of 30°N–25°S) are damped persistence of the mean anomalies observed the previous month (added to the forecast season's climatology), with an *e*-folding time of 3 months (Mason et al. 1999). In the tropics, multimodel, mainly dynamical SST forecasts are used for the Pacific—the basin having the best known physics and model forecast consistency—while statistical and dynamical forecasts are combined for the Indian and Atlantic Oceans. In the non-Pacific tropical basins, during seasons having little apparent SST predictive skill, damped persistence of the SST anomalies observed in the most recent month are used, but with a lower damping rate than applied in the extratropics. For seasons having greater apparent skill, canonical correlation analysis (CCA; Glahn 1968; Barnett and Preisendorfer 1987) models are used in the Indian Ocean (Mason et al. 1999) and tropical Atlantic Ocean (Repelli and Nobre 2004).

A separate scenario of globally persisted SST anomaly, consisting of undamped anomalous SST observed the previous month added to the climatology of the months being forecast, is used out to 4 months for the IRI's shortest lead time forecasts. For the nonpersisted, evolving SST anomaly predictions, the AGCMs are run out to 7 months.

The methods used to develop the SST forecasts are detailed in Table 1. For the evolving SST forecasts, three versions of the forecast SST anomalies have been used. In the first version, used through May 2004, a single deemed best estimated forecast SST scenario was used for tropical Pacific SST, which was that of the National Centers for Environmental Prediction (NCEP) coupled model (Ji et al. 1998). Beginning June 2004, three separate tropical Pacific scenarios were used: NCEP's more recently developed global coupled model [Climate Forecast System (CFS); Saha et al. 2006], the Lamont-Doherty Earth Observatory (LDEO) intermediate coupled model version 5 (Chen et al. 2004), and the constructed analog (CA) statistical model (Van den Dool 1994, 2007; Van den Dool et al. 2003). One-third of the ensemble members of each of the AGCMS was forced by each SST scenario. This multiscenario design (Li et al. 2008) was believed to better represent the uncertainty expressed by the spread of the ensemble mean SST forecasts among the three models, whose forecast ENSO states often differed considerably. In the tropical Atlantic and Indian Oceans, a single scenario was used, consisting of the average of the CFS and CA ensemble mean forecasts.

Use of multiple SST scenarios was refined further in a third version starting in May 2007, noting that sometimes the ensemble mean tropical Pacific forecasts of the three models agreed closely, while at other times they differed greatly. The degree of disagreement is not believed to be significantly related to actual forecast uncertainty (e.g., Kharin and Zwiers 2002; Tippett et al. 2007). To ensure more approximately comparable scenario differences from year to year for the same forecast start month and lead time, the three scenarios were derived based on the historical error of the 3-way superensemble mean of the models, for hindcasts using observed SSTs over *the global tropics*. The preferred structures of the error field were found using principal components analysis (PCA) on the multimodel mean SST hindcast error. The three scenarios then used are 1) the 3-way multimodel ensemble mean SST forecast itself (with mean biases removed, and that mean 2) plus and 3) minus the first PC of the historical error. The PC accounts for roughly 40% of the model error variance, and its spatial pattern for most start and lead times is related largely, but not exclusively, to ENSO.

b. AGCMs for climate prediction

In the second tier of IRI's prediction system, several AGCMs are forced by the set of predicted SSTs. The initial states of the AGCMs are not based on observed atmospheric or land surface conditions but are taken from ongoing updates to long AGCM simulations forced

TABLE 1. Versions of SST forecast used for each tropical ocean basin by IRI to force its multiple AGCMs. The persisted SST (first row) is used for entire globe for the first 3-month forecast season only, while the evolving SST forecasts are used for forecasts for all lead times.

SST version	Period of use	Ensemble apportionment	Tropical Pacific SST	Indian Ocean SST	Tropical Atlantic SST
Persisted SST	Oct 1997–present	Uniform	Undamped persisted	Undamped persisted	Undamped persisted
Evolving SST 1	Oct 1997–May 2004	Uniform	1) NCEP coupled[a]	CCA	CCA or damped persisted
Evolving SST 2	Jun 2004–Apr 2007	One-third for each scenario	Separately: 1) CFS[b] only 2) CA[c] only 3) LDEO[d] only	Mean of CFS[b] and CA[c]	Mean of CFS[b] and CA[c]
Evolving SST 3	May 2007–present	One-third for each scenario	Separately: 1) Mean of CFS, CA, LDEO 2) Same as 1, plus perturbation[e] 3) Same as 1, minus perturbation	Separately: 1) Mean of CFS, CA, CCA 2) Same as 1, plus perturbation[e] 3) Same as 1, minus perturbation	Separately: 1) Mean of CFS, CA 2) Same as 1, plus perturbation[e] 3) Same as 1, minus perturbation

[a] NCEP coupled ENSO forecast model (Ji et al. 1998).
[b] NCEP Climate Forecast System (Saha et al. 2006).
[c] Constructed analog statistical model (Van den Dool 1994, 2007; Van den Dool et al. 2003).
[d] Lamont-Doherty Earth Observatory intermediate coupled model 5 (Chen et al. 2004).
[e] Perturbation consists of first EOF of historical error of mean of CFS, LDEO, and CA.

by observed SSTs. Because the earliest predicted period begins 3–4 weeks after the time of the forecast integrations, use of observed atmospheric initial conditions is not considered critical. However, the lack of observed land surface initial conditions (soil moisture, snow cover) may slightly degrade the forecasts because their effects can continue for longer than one month. The initial conditions used, differing among ensemble members, are characteristic of the respective model, region, and time of year, and the probability distribution of possible atmospheric states is spanned across members, constrained to be consistent only with the prescribed SST boundary conditions.

The number and specific set of AGCMs, and their forcing by the SST predictions, have evolved over the 11 yr of forecasting (Table 2). Three AGCMs with T42 spectral horizontal resolution (~2.8° latitude–longitude) were used from late 1997 to early 2001, after which additional or replacement AGCMs were used (Barnston et al. 2003). Seven AGCMs have been used from late 2004 through 2008, providing a total of 144 (68) ensemble members forced by evolving (persisted) SST. The National Aeronautics and Space Administration (NASA), Center for Ocean–Land–Atmosphere Studies (COLA), Geophysical Fluid Dynamics Laboratory (GFDL), and Scripps models have highest horizontal resolution [T62 spectral (~2.0°) or 2.5° × 2.0° gridded]. The European Center for Medium-Range Weather Forecasts–Deutsches Klimarechenzentrum: Hamburg Model (ECHAM) and National Center for Atmospheric Research (NCAR) Community Climate Model (CCM) have been run at IRI, while the other models have been run at their home institutions using IRI's SST boundary conditions and graciously sent monthly to IRI to contribute to the forecasts. All model outputs are expressed with respect to their own climatologies (e.g., mean and terciles) based on multidecadal simulations using observed SST.

The climatological base period used as the reference frame for forecasts and observations was 1961–90 from 1997 until June 2001, 1969–98[2] from July 2001 through 2002, and 1971–2000 from January 2003 to present.

Forecasts issued through early 2001 were developed largely from the ECHAM3.6, CCM3.2, and NCEP–Medium-Range Forecast (MRF9) AGCMs, whose forecasts were combined subjectively by the forecasters using various model validation statistics (Mason et al. 1999; Goddard et al. 2003). Forecast formation was further guided by empirical probabilistic composites based on relative frequencies of occurrence of tercile-based categories keyed to past ENSO episodes (Mason and Goddard 2001). Beginning in mid-2001 the process of merging the AGCM predictions into a final forecast was automated

[2] This nonstandard base period was used because of delays in updating of the global climate observational data used.

TABLE 2. AGCMs used to develop IRI's climate forecasts: basic features and references. ECPC is the Experimental Climate Prediction Center, and AM2 denotes the Atmospheric Model, version 2.

AGCM	Horizontal resolution	Vertical resolution (layers)	Period of contribution to IRI forecasts	No. of ensemble members (evolving/persisted SST)	Model development site
CCM3.2[a]	T42	18	Sep 97–Aug 03	10/10	NCAR[b]
CCM3.6[a]	T42	18	Dec 04–present	24/24	NCAR[b]
COLA	T63	13	Oct 01–May 04 Jun 04–present	10/0 12/0	COLA[c]
ECHAM3.6[a]	T42	19	Sep 97–Jan 02	10/10	Max Planck Institute[d]
ECHAM4.5[a]	T42	19	Aug 01–present	24/24	Max Planck Institute[e]
ECPC Noah	T62	18	Jun 03–May 04 Jun 04–present	10/10 12/10	Scripps Institution of Oceanography[f]
GFDL AM2p12b	2.5° × 2.0°	18 or 24	Oct 04–present	30/10	GFDL[g]
NASA Seasonal-to-Interannual Prediction Project (NSIPP)	2.5° × 2.0°	34	Apr 01–May 04 Jun 04–present	9/0 12/0	Goddard Space Flight Center (GSFC)[h]
NCEP MRF9	T40	18	Sep 97–May 04 Jun 04–present	10/0 30/0	NCEP[i]

[a] Model is run at IRI.
[b] Hack et al. (1998); Hurrell et al. (1998); Kiehl et al. (1998).
[c] Schneider (2002).
[d] Deutsches Klimarechenzentrum (1992); Roeckner et al. (1992).
[e] Roeckner et al. (1996).
[f] Kanamitsu et al. (2002); Kanamitsu and Mo (2003).
[g] GFDL Global Atmospheric Model Development Group (2004).
[h] Bacmeister et al. (2000); Pegion et al. (2000); Schubert et al. (2002).
[i] Kumar et al. (1996); Ji et al. (1998); Livezey et al. (1996).

(Barnston et al. 2003). Two multimodel ensembling methods were used: a Bayesian method (Rajagopalan et al. 2002; Robertson et al. 2004) and a canonical variate method (Mason and Mimmack 2002), and the two forecast results were averaged. In both methods, individual model weighting varies by grid point and forecast target season, governed by the models' historical skills over an approximately 50-yr period when forced by observed SST fields. Use of this model weighting formulation, to be discussed in the context of the skill results in section 4f, is not ideal because observed SST is not available for the target periods in the real-time forecast setting.

c. Final forecast

Even with the more automated system implemented in 2001, final minor subjective modification of the objective forecasts by the forecasters has continued. This modification has consisted largely of overall damping of probabilities toward climatology—more at high than at low latitudes, and in particular for inordinately strong regional probability shifts. Light spatial smoothing has also been done to reduce noise. Other modifications include selected spatial model output statistics (MOS) corrections of systematic errors of the individual AGCMs for precipitation for specific regions using CCA (Ndiaye et al. 2009; Tippett et al. 2003; Landman and Goddard 2002); a nudging toward reduction (enhancement) of probabilities for below-normal (above normal) temperature, partly in response to a diagnostic verification of IRI's forecasts during 1997–2000 (Wilks and Godfrey 2002); and making the forecasts more consistent with those of other meteorological centers or regional climate outlook forums.

3. Data and methods

a. Data

Consistent datasets of observed global temperature and precipitation are required to calibrate the model forecasts and to verify the forecasts. For temperature, the 2° gridded global Climate Anomaly Monitoring System (CAMS) dataset from National Oceanic and Atmospheric

TABLE 3. Information and computation for RPSS, LSS, and GROC verification measures.

Verification measure; conceptual basis; references	Computation formula
Ranked probability skill score (RPSS): Squared probability error, cumulative across forecast categories, between forecast probability and observed "probability" (0 if not observed, 1 if observed), compared with same calculation for a naive reference forecast such as that for climatological probabilities; highly analogous to mean squared error skill score for deterministic forecasts (Epstein 1969; Wilks 2006).	RPSS = $1 - (\mathrm{RPS}_{\mathrm{fct}}/\mathrm{RPS}_{\mathrm{ref}})$, where fct refers to the forecasts, and ref refers to a naive reference forecast such as 1/3 for each tercile category. Here RPS = $(1/\mathrm{nfct})\sum_{\mathrm{ifct}=1}^{\mathrm{nfct}}\left[\sum_{\mathrm{icat}=1}^{\mathrm{ncat}}(\mathrm{PCUMfct}_{\mathrm{icat}} - \mathrm{PCUMobs}_{\mathrm{icat}})^2\right]$, where icat is category number, ncat is the total number of categories, ifct is forecast number, and nfct is the total number of forecasts.
Likelihood skill score (LSS): Geometric mean of the probability assigned to the category that was actually observed (Aldrich 1997).	LSS = $(\mathrm{LIK}_{\mathrm{fct}} - \mathrm{LIK}_{\mathrm{ref}})/(1 - \mathrm{LIK}_{\mathrm{ref}})$, where fct and ref are as defined for RPSS. Here LIK = $\sqrt[\mathrm{nfct}]{\pi P_{\mathrm{ifct}}}$, where nfct is the number of forecasts; π is the multiplication operator with range ifct = 1 to nfct.
Generalized relative operating characteristics (GROC): Proportion of all available pairs of observations of differing category whose probability forecasts are discriminated in the correct direction; equivalent to ROC area for individual categories, and somewhat analogous to Spearman rank correlation, and other rank tests, for ranked deterministic forecasts (Mason and Weigel 2009).	GROC = $\left[\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} I(p_{k,i}, p_{l,j})\right] \Big/ \left(\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l\right)$, where $I(p_{k,i}, p_{l,j}) = 1$, 0.5, or 0, depending on whether difference in the probability forecasts was in correct direction, was neutral, or was in the incorrect direction, respectively [see Mason and Weigel (2009) for formula for determining which $I(p_{k,i}, p_{l,j})$ outcome occurred]. Here m_v is the number of categories, k and l identify the differing observed categories whose forecasts are compared, and i and j are forecast numbers within the specified categories. The numerator sums the $I(p_{k,i}, p_{l,j})$ outcomes over all qualifying pairs of forecasts; the denominator contains the number of pairs.

Administration (NOAA) (Ropelewski et al. 1985) is used. For precipitation, the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP; Xie and Arkin 1997) for data from 1979 onward and the 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. Tests for relative biases during the overlap period indicate minor biases in mean and biases in variance (CRU data having lower variance). The latter biases slightly affect the terciles when the 1961–90 climatology period was used but have little effect for the two later base periods.

b. Methods

Here we use the ranked probability skill score (RPSS; Epstein 1969), a likelihood score (Aldrich 1997), and a generalization of the relative operating characteristics (ROC) curve (Mason 1982) to the three forecast categories collectively (Mason and Weigel 2009). For additional diagnostic understanding, we apply reliability analysis (Murphy 1973).

The RPSS (Epstein 1969; Wilks 2006), an extension of the Brier skill score (Brier 1950) to more than two categories, begins with computation of the ranked probability score (RPS). RPS is the squared probability error, cumulative across the forecast categories in ascending rank order, between the categorical forecast probability and the corresponding observed "probability" (100% probability assigned to the observed category, 0% otherwise). Higher RPS indicates larger forecast error. RPSS is positive when the RPS of the forecasts is less than that of a chosen reference forecast (here, the climatology forecast of one-third probability for each category).

The likelihood score (Aldrich 1997) is based on the product, over all forecasts in a set of n forecasts, of the probabilities forecast for the actually observed category. The nth root of this product is taken, yielding an intuitively meaningful geometric mean probability assigned to the correct category. The likelihood score is closely related to the ignorance score (Roulston and Smith 2002), linked to information theory, and derived scores such as return ratio (Hagedorn and Smith 2008; Tippett and Barnston 2008). A likelihood skill score (LSS) compares the likelihood score for the forecast with that of a reference forecast (here, climatology), assigning zero skill if it equals the reference. A differing feature between LSS and RPSS is that LSS is based on the probability assigned only to the category later observed (the *locality* property; Brocker and Smith 2007), ignoring probabilities for the other categories; RPSS uses the probabilities forecast for all three categories and gives greater credit when high probabilities are assigned to a category adjacent to that observed versus a more distant category. Both RPSS and LSS are used to verify the IRI forecasts in part to assess the extent to which the simpler LSS provides information about forecast quality

similar to (or as fully as) the more comprehensive and widely used RPSS.

Another probabilistic verification measure is an extension of the ROC area to include all forecast categories collectively (Mason and Weigel 2009). This generalized ROC score (GROC) is the proportion of all available pairs of observations of differing categories whose probability forecasts are discriminated in the correct direction. With a possible range of 0%–100%, a 50% rate of correct discrimination is expected by chance. The calculations of RPSS, LSS, and GROC are summarized in Table 3. RPSS and LSS can be calculated for individual forecasts, permitting a time series of forecast skills. By contrast, GROC is calculated only for a set of forecasts, of which at least two must have differing observational outcomes.

All three of RPSS, LSS, and GROC are *proper* (Winkler and Murphy 1968; Brocker and Smith 2007)—that is, they cannot be enhanced by making the forecast probabilities different from those believed true by the forecasters (''hedging''; Murphy and Epstein 1967). A difference between GROC on the one hand, and RPSS and LSS on the other, is that the goodness of the calibration of the probabilities matters to the latter two scores, while it is essentially irrelevant to GROC. GROC evaluates purely discrimination ability within the forecast sample at hand, without penalty for overall or conditional biases in the probability values. The other side of the same coin is that GROC does not reward correct forecasts of overall shifts of climate in the forecast sample relative to a longer period of reference such as a warmed climate relative to a past 30-yr period. The above characteristics are similar to those of the temporal correlation coefficient for verification of deterministic forecasts and in contrast to the mean squared error or Heidke skill scores (Barnston 1992)—the latter two, like RPSS and LSS, being calibration sensitive.

Limited ensemble sizes restrict forecast probabilities to finite numbers of possible values, creating small offsets from asymptotic probabilities coming from theoretically infinite ensemble sizes. These offsets slightly decrease RPSS (Weigel et al. 2007a,b), whose climatology reference RPS remains ''perfect.'' Impacts on LSS and GROC are smaller. Adjustments for these biases are not conducted here, given IRI's fairly large ensemble size.

In addition to the above verification measures, the probabilistic reliability of the forecasts is diagnosed using attributes diagrams (Murphy 1973; Wilks 2006). These show the correspondence of the full range of issued forecast probabilities and their associated relative frequency of observed occurrence, revealing forecast characteristics such as probabilistic bias, forecast over-(under-) confidence, and forecast sharpness.

4. Results

The temporal variability of the performance of IRI's forecasts is shown by time series of a verification measure (RPSS or LSS) for a given lead time averaged over the globe, the tropics, or specific regions. These scores are averaged over the scores for each grid square, area weighted by the cosine of their latitude. Additionally, the geographical distribution of the forecast quality is shown by computing the measures for each grid square over all forecasts or a subset of forecasts (e.g., for a given season and/or lead time). We show the performance of the issued forecasts as well as the objective multi-AGCM output used as the primary guiding tool. First, however, we consider the quality of the SST forecasts.

a. SST forecast skill

Favorable performance of the climate forecasts in a two-tiered design depends on performance in critical aspects of the SST forecasts—particularly the ENSO state and the SST anomaly patterns in the tropical Atlantic and Indian Oceans. Figure 1 shows the spatial distribution of temporal correlation between tropical SST forecasts and observations at 0.5- and 3.5-month lead times, and Fig. 2 shows time series of the forecasts and observations averaged over several key rectangular areas. Figure 2 (top) shows the performance of the SST forecasts in capturing the ENSO state as represented by the Niño-3.4 SST index (Barnston et al. 1997) at 0.5- and 3.5-month lead times. Forecasts and observations correlate 0.88 and 0.75 at 0.5- and 3.5-month lead times, respectively, indicating useful skill in anticipating the ENSO-related SST. Omitting the strong El Niño through the first half of 1998, these correlations drop to 0.86 and 0.73, suggesting that the skill did not depend heavily on this one episode.

SST forecasts in the Indian and tropical Atlantic Oceans (Figs. 1 and 2) were comparatively less skillful, and skills differ little from persistence-based forecasts (Table 4). These lower skills are consistent with the weaker inherent predictability of SST in the non-Pacific tropical ocean basins (Goddard et al. 2001; Stockdale et al. 2006). Although interannual variability of tropical SSTs outside of the central and eastern Pacific is small (Table 4), anomaly patterns in these oceanic regions are believed key to enhanced likelihoods for specific climate anomalies (e.g., Chang et al. 2006). For example, in parts of tropical and subtropical Africa, Asia, and South America, climate anomalies are related to a zonal dipole in the Indian Ocean (Saji et al. 1999; Goddard and Graham 1999), an El Niño–like structure in the equatorial Atlantic (Zebiak 1993), and meridional gradients in the tropical Atlantic (Ward and Folland 1991;

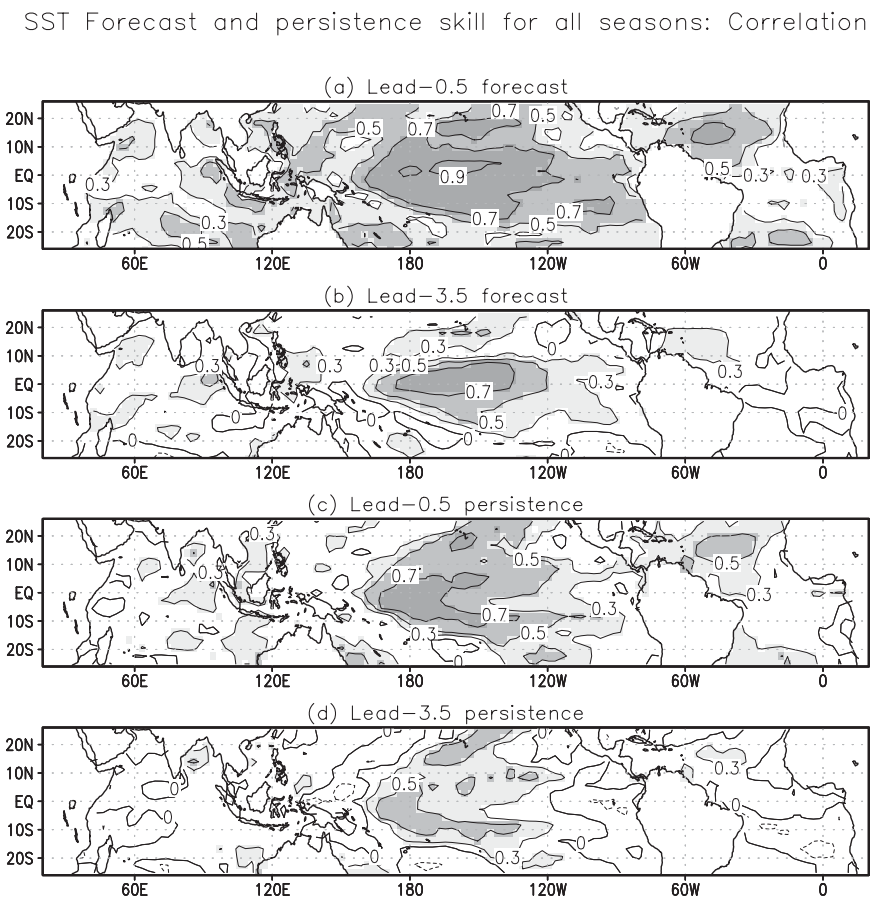

FIG. 1. Spatial distribution of correlation coefficient between seasonal mean tropical SST forecasts and observations for all seasons at (a) 0.5- and (b) 3.5-month lead times, and likewise for persistence forecasts of 3-month mean SST anomaly at (c) 0.5- and (d) 3.5-month lead times.

Enfield et al. 1999; Servain et al. 1999). Both tropical Indian and Atlantic Ocean SSTs appear sensitive to exogenous, and sometimes extratropical, phenomena that may have little inherent predictability (Kushnir et al. 2006). While this may be true for the tropical Pacific as well, the Pacific has better defined, slower, and stronger internal dynamics that frequently outweighs exogenous influences.

b. Temporal variability of climate forecast skill

Figure 3 shows time series of RPSS averaged over the near-global and tropical (25°N–25°S) land areas for forecasts for each of the four lead times (0.5, 1.5, 2.5, and 3.5 months) from the period October–December (OND) 1997 to December–February (DJF) 2008/09 for precipitation and temperature. Forecasts of climatological probabilities are included. The proportions of land area coverage by nonclimatology forecasts for the globe, tropics, and extratropics (Table 5) indicate highest proportions of nonclimatology forecasts issued for the tropics, for temperature, and for shorter lead times. Nonclimatology forecasts are somewhat more prevalent in forecasts for which ENSO extremes were expected than otherwise and for boreal autumn and winter than other seasons because of greater confidence in the forecast ENSO state for those seasons.

Forecast skill over the 11-yr period has been strongly related to ENSO variability (Fig. 3). Correlations between the absolute value of the Niño-3.4 SST anomaly and tropical RPSS for precipitation are 0.54, 0.44, 0.40, and 0.43 for 0.5-, 1.5-, 2.5-, and 3.5-month lead precipitation forecasts, respectively. (Corresponding Spearman rank correlations are 0.44, 0.44, 0.41, and 0.38.) Figure 4 (left) shows the effect of the ENSO state on RPSS for 0.5-month lead tropical precipitation forecasts as a function of lag time between the season of the ENSO state and that of the climate forecast target. Despite modest average skill levels, a simultaneous positive relationship with both phases of ENSO is noted (consistent with results in Goddard and Dilley 2005), El Niño being associated with greater skill than La Niña. Figures 3 and 4 show near-zero precipitation skills during ENSO-neutral

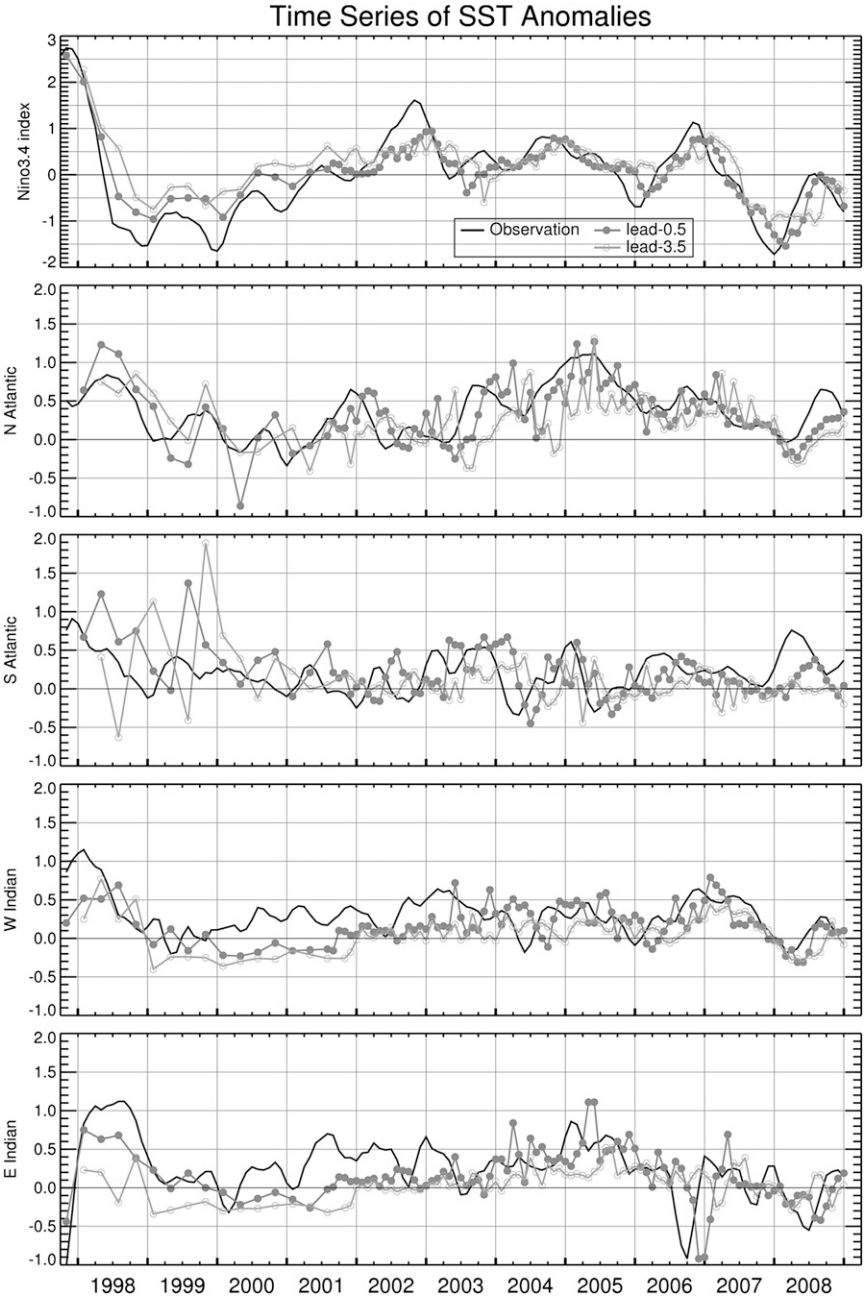

FIG. 2. IRI's predictions of SST anomaly from late 1997 through 2008 in five regions at 0.5- and 3.5-month leads, together with the corresponding observations: Niño-3.4 (east-central tropical Pacific), north tropical Atlantic, south tropical Atlantic, west Indian Ocean, and southeast Indian Ocean. The center month of the 3-month period being forecast is indicated on the horizontal axis. Vertical year-separating lines are drawn through the DJF season. Regional lat–lon boundaries and the correlation skills of these forecasts and of corresponding persistence forecasts are provided in Table 4.

periods in both extratropics and tropics, which is comparable to Livezey and Timofeyeva (2008), who identified ENSO variability as virtually the sole source of seasonal precipitation forecast skill for the United States.

The time series of RPSS for IRI's temperature forecasts (Fig. 3) show higher average levels than those of precipitation forecasts. Temperature skill is related to ENSO state but differently than precipitation: skill is

TABLE 4. Skill (as correlation coefficient), bias, and variance ratio with respect to observations of IRI's predictions of SST (1997–2008) in five rectangular tropical ocean regions at 0.5- and 3.5-month lead times. Skill of forecasts of simple persistence of observed seasonal anomalies at the same lead times is shown for comparison. Boldface correlations attain statistical significance at 95% level using 1 degree of freedom per year for the Niño-3.4 region, and 2 degrees of freedom per year for the other SST regions.

SST region	Location	Lead time (months)	No. forecasts	Correlation (forecast/persistence)	Avg obs anomaly (°C)	Bias (°C)	Obs std dev (°C)	Variance ratio
Niño-3.4	5°N–5°S, 120°–170°W	0.5	104	**0.88/0.62**	0.08	−0.03	0.80	0.66
		3.5	101	**0.75**/0.24	0.06	0.04	0.76	0.66
West tropical Indian	10°N–10°S, 50°–70°E	0.5	104	0.36/0.39	0.29	−0.11	0.23	1.07
		3.5	101	**0.47**/0.08	0.29	−0.22	0.23	0.83
Southeast tropical Indian	0°–10°S, 90°–110°E	0.5	104	**0.44**/0.33	0.24	−0.08	0.36	0.80
		3.5	101	0.21/0.21	0.25	−0.20	0.35	0.31
North tropical Atlantic	20°–5°N, 30°–60°W	0.5	103	**0.69/0.61**	0.40	−0.06	0.32	1.29
		3.5	100	0.27/0.29	0.40	−0.17	0.33	0.88
South tropical Atlantic	0°–20°S, 30°W–10°E	0.5	103	0.18/0.11	0.19	0.00	0.25	1.43
		3.5	100	−0.10/−0.14	0.19	−0.11	0.25	1.44

highest near the end of, and shortly following, El Niño events, and lowest with the same timing for La Niña events. Figure 4 (right) shows the effect of the ENSO state on RPSS for 0.5-month lead tropical temperature forecasts as a function of lag time between the SST and the climate forecast target. The greatest impact of ENSO on RPSS occurs 4 months following the ENSO peak for both ENSO phases. This influence on forecast skill is attributable to a delayed temperature response in both tropics and extratropics (Kumar and Hoerling 2003), which was earlier documented in the context of the atmospheric bridge (Lau and Nath 1996; Alexander et al. 2002) and strongly exemplified in the response to the 1997/98 El Niño (Kumar et al. 2001).

One reason for the comparatively higher overall temperature forecast skill is that the skill receives a substantial contribution from correctly forecasting increased probabilities of above-normal temperature related to global warming. This warming is partially reproduced by the AGCMs, forced by SSTs that reflect part of the global warming signal, although the climate change signal is largely lost in the SST forecasts, and consequently in the AGCM responses, after the first few months in models using fixed (and now outdated) greenhouse gas settings (Doblas-Reyes et al. 2006; Liniger et al. 2007). Therefore, the warming signal is further captured by the forecasters who make additional subjective probabilistic adjustments toward warmth. The climate change component of skill is much weaker for precipitation, whose trends are generally smaller and may be of either sign, depending on location and season. Seasonal temperature is subject to well established probabilistic shifts related to ENSO (e.g., Halpert and Ropelewski 1992), providing a source of interannual predictability largely independent of the warming trend. Although the geographical distribution of ENSO's effects on temperature differs from that associated with global warming, there are similarities between effects of El Niño and global warming, particularly in the tropics. Consequently, in a tropical average sense, El Niño tends to amplify the effects of global warming, yielding increased confidence in forecasts of above-normal temperature, while La Niña tends to weaken or cancel global warming effects, resulting in a smaller net signal, greater forecast uncertainty, and lower skill.

These ideas appear substantiated by Figs. 3 and 4, showing highest (lowest) temperature skills during and after El Niño (La Niña) events, particularly in the tropics. Correlations between the Niño-3.4 SST anomaly and the tropical RPSS 4 months later are high: 0.80, 0.77, 0.76, and 0.72 for 0.5-, 1.5-, 2.5-, and 3.5-month lead time forecasts, respectively. (Corresponding Spearman rank correlations are 0.79, 0.78, 0.78, and 0.74.) In their evaluation of IRI's forecasts during 1997–2001, Goddard et al. (2003) concluded that empirical ENSO probabilistic composites were not helpful for IRI's seasonal temperature forecasts because the La Niña conditions during the majority of the period led to increased forecast probabilities for below-normal temperature, while above-normal temperatures continued to predominate in the observations.

c. Seasonality and geographical distribution of climate forecast skill

Figure 5 shows the geographical distribution of RPSS over the globe for all seasons for precipitation and temperature at 0.5-month and 3.5-month lead times.

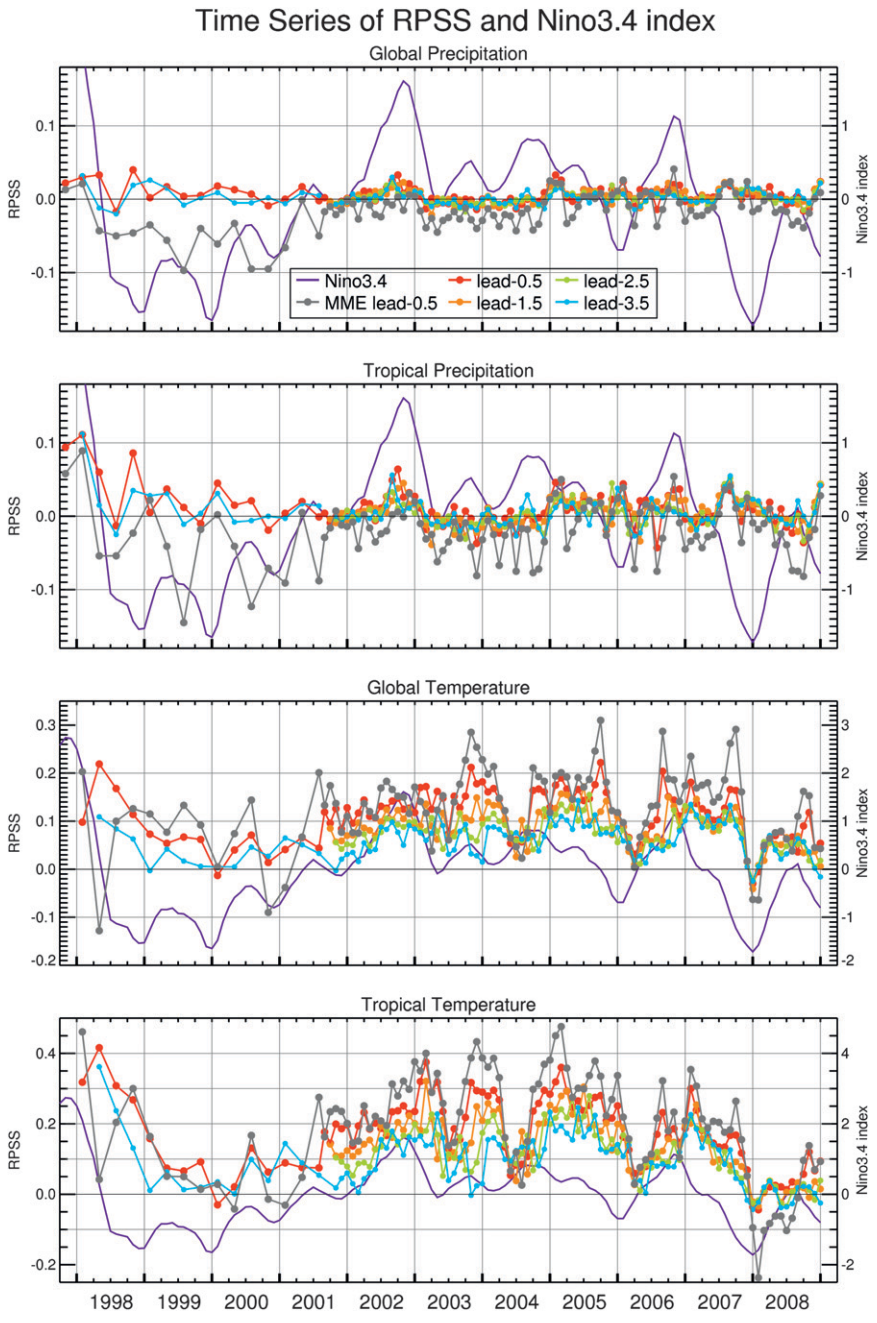

FIG. 3. Time series of RPSS averaged over the tropical land areas (25°N–25°S) or for the globe (except Antarctica) for forecasts for each of the four lead times (0.5, 1.5, 2.5, and 3.5 months) from late 1997 through 2008 for (a) global precipitation, (b) tropical precipitation, (c) global temperature, and (d) tropical temperature. The RPSS of the objective multimodel ensemble forecasts, used as essential guidance for the final forecasts, is shown for 0.5-month lead. The purple curve shows the Niño-3.4 SST observation. The center month of the 3-month period being forecast is indicated on the horizontal axis. Vertical year-separating lines are drawn through the DJF season. Note differing ordinate scales across the panels.

TABLE 5. Percentage of land area for which nonclimatology forecasts were issued, 1997–2008.

	Precipitation			Temperature		
Forecast lead time (months)	Globe (%)	Tropics (%)	Extratropics (%)	Globe (%)	Tropics (%)	Extratropics (%)
0.5	28	44	16	64	77	54
1.5	23	39	11	56	68	47
2.5	21	36	10	51	63	42
3.5	20	34	10	46	58	37

Relatively high temperature skill is noted in much of the tropics and in some extratropical regions. Temperature skill decreases, but does not disappear, with lead time. Skill for precipitation is lower than that for temperature but is also generally highest in the tropics. While precipitation skill averaged over all seasons does not disappear at 3.5-month lead, it decays more quickly with lead time than temperature, proportionally with respect to its initial level, in regions having highest season-specific skills (not shown). Skill is generally greater for temperature than for precipitation partly because of the more pervasive and unidirectional manifestation of climate change (and generally correct forecasts for such) in temperature than precipitation. To first order, the global warming component of temperature forecast skill pervades all seasons, all lead times, and most regions and is proportionately most prominent in the tropics where interannual and internal variability are generally weakest. (Warming is greater in the extratropics in degrees Celsius but is outweighed by still greater amounts of interannual variability.)

Because RPSS for precipitation is below −0.01 over nearly as much area of the globe as it is above 0.01 at 0.5 and 3.5-month leads (Fig. 5), one reasonably might question the field significance of the skill result (Livezey and Chen 1983). Monte Carlo tests were conducted in which the years were shuffled 5000 times, while the ordering of the months within a given year remained intact to represent the effective sampling time for an ENSO cycle. The global mean RPSS for the shuffled data never attained the level of the actual verification (at 0.006) at 0.5-month lead and exceeded it (at 0.003) in 1 out of 5000 trials at 3.5-month lead. Because field significance is strong, one can trust not only the existence of real global mean skill but also the general features of the skill's geographical distribution.

The geographical distribution of mean RPSS for precipitation forecasts at 0.5-month lead is shown in Fig. 6

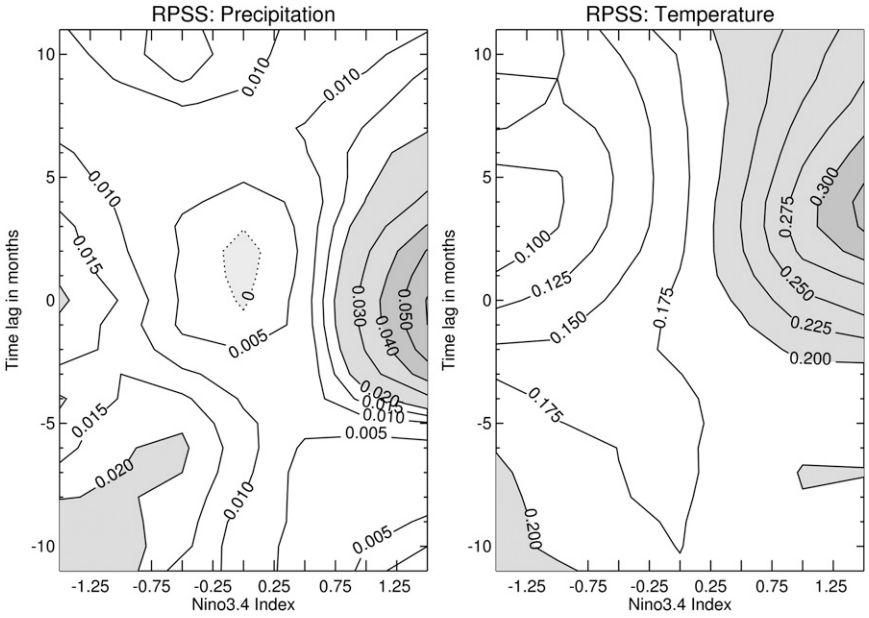

FIG. 4. Contoured plot of RPSS for tropical (a) precipitation and (b) temperature forecasts at 0.5-month lead as a function of the ENSO state (represented by Niño-3.4 SST anomaly) and the time lag (months) between the SST and the forecast climate (temperature or precipitation). The lag is positive when SST precedes the time of the climate forecast. Variable contour intervals and shading thresholds are arbitrary, for readability.

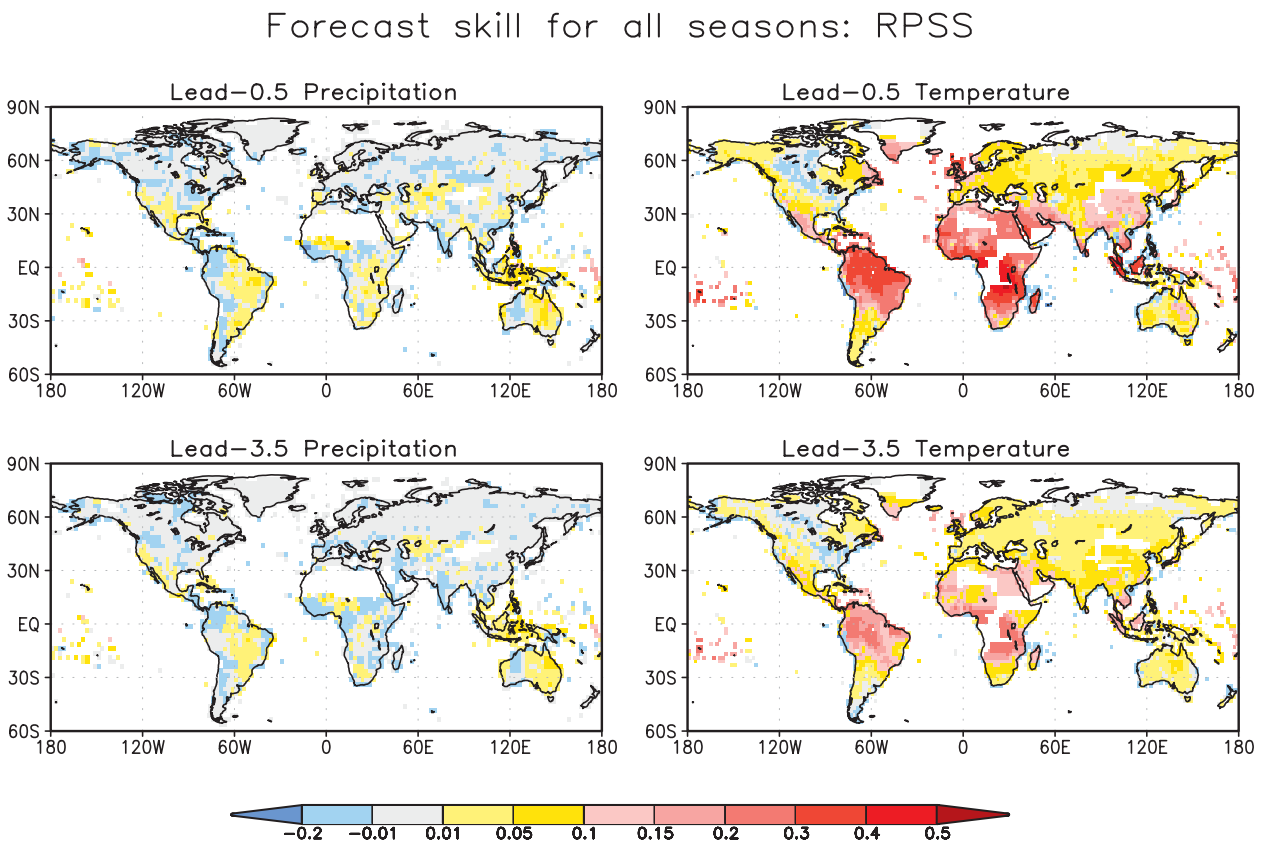

FIG. 5. Geographical distribution of RPSS averaged over all seasons for (left) precipitation at 0.5-month lead and 3.5-month lead and (right) likewise for temperature. White areas lack sufficient data to calculate RPSS meaningfully.

for January–March (JFM), April–June (AMJ), July–September (JAS), and OND. Skills are often related to the seasonal cycle of rainfall itself, which in the tropics–subtropics maximizes with the local summer monsoon season or with the twice-yearly passage of the ITCZ near the equator. Skill is highest in Indonesia, eastern equatorial Africa, and southeastern South America during the last few months of the calendar year; in portions of southern Africa from November to March; and in India and the Sahel from June through September. Skill is concentrated in the seasons and regions having known responses to ENSO (Ropelewski and Halpert 1987, 1989, 1996; Mason and Goddard 2001) as well as some additional areas in response to SST anomalies outside the tropical Pacific (e.g., the African Sahel and Guinea coast, and Northeast Brazil in response to the tropical Atlantic).

The seasonal cycles of precipitation forecast skill for several regions having well-defined monsoon seasons and/or ENSO-related responses are shown for all 12 running 3-month seasons in Fig. 7, using RPSS and GROC. Skill in the Philippines is highest in late boreal winter following the usual peaking of ENSO episodes and minimal in boreal summer during the southwest Asian monsoon. Indonesia and the western tropical Pacific islands show maximum skill in late boreal autumn, when ENSO episodes often mature and the ITCZ migrates through from north to south. Skill in the African Sahel peaks during the late boreal summer rainy season, and analogous behavior holds for austral summer in southern Africa. In eastern equatorial Africa, skill peaks in late boreal autumn (short rainy season) but is low during the boreal spring (long rainy season), as established through ENSO responses empirically (Ropelewski and Halpert 1987; Mason and Goddard 2001) and physically in the context of the intermediary role of the Indian Ocean SST in the short rainy season (Goddard and Graham 1999). In the southern United States, Mexico, and the Caribbean, skills peak during boreal winter, during a dry season, following known teleconnections to ENSO (Ropelewski and Halpert 1987, 1989; Mason and Goddard 2001).

While there are no large differences between the skill pictures painted by RPSS and GROC beyond their differing scaling, a tendency for fewer cases of negative skill is noted in GROC (<0.5) than in RPSS (e.g., during boreal summer in western tropical Pacific islands, RPSS < 0 while GROC > 0.5). This is likely due to the

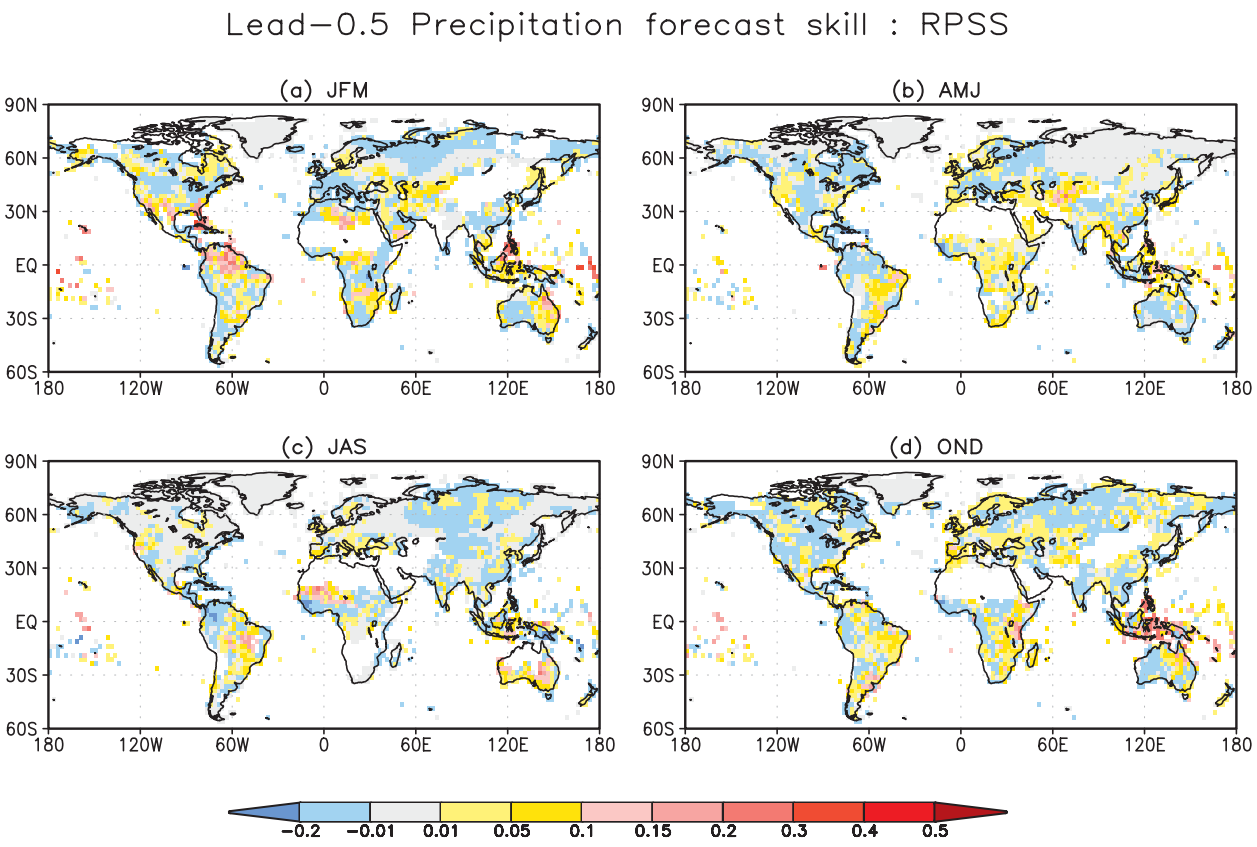

FIG. 6. Geographical distribution of mean RPSS for precipitation for JFM, AMJ, JAS, and OND seasons at 0.5-month lead time. White areas lack sufficient data to calculate RPSS meaningfully.

presence of discrimination in the forecasts, such as cases of above-normal rainfall being forecast with higher probabilities of above normal than cases lacking above-normal rainfall, even if the probability values have systematic biases. Unconditional or conditional (rainfall dependent) forecast biases, such as most probabilities for above normal being too low, would be penalized in RPSS, counteracting credit given for the probability for above normal being relatively higher for cases of observed above-normal rainfall than for other cases; GROC would reflect such discriminative ability, unhidden by biases. Hence, Fig. 7 tells us that the precipitation forecast probabilities may not have been optimally calibrated; this will be examined below in a reliability analysis.

The geographical distribution of mean RPSS for temperature forecasts at 0.5-month lead time are shown in Fig. 8 for JFM, AMJ, JAS, and OND. The skill patterns of the four seasons do not differ greatly. The JFM season features the greatest spatial extremes of skill, with highest tropical skill and most widespread negative extratropical skill. The seasonal cycle of temperature forecast skill for selected large regions (Fig. 9) shows smaller, subtler seasonal dependence than that of precipitation skill. This difference likely exists because much of the temperature skill is related to correctly forecast probability shifts toward above normal because of global warming, which is largely independent of season. Temperature also lacks the tropical seasonal migratory cycles found in precipitation (e.g., ITCZ, monsoons). Skill in many regions (e.g., northern South America, Indonesia) is slightly higher in boreal winter than summer, which is likely related to seasonality of the ENSO cycle: El Niño (La Niña) tends to warm (cool) the tropical atmosphere most predictably and strongly near and following its mature phase late in the calendar year.

The differences in skill shown by RPSS and GROC (Fig. 9) indicate fewer cases of negative RPSS than GROC (<0.5), for example, late boreal summer skills for Africa. This difference is due to the credit given by RPSS for correctly elevated probabilities for above-normal temperature caused by global warming, even in the absence of correct year-to-year discrimination among probability values within the 11-yr period. Because of the reference forecast used in RPSS (a climatology based on a completed 30-yr period), forecasts uniformly tending toward above-normal temperature earn credit in RPSS

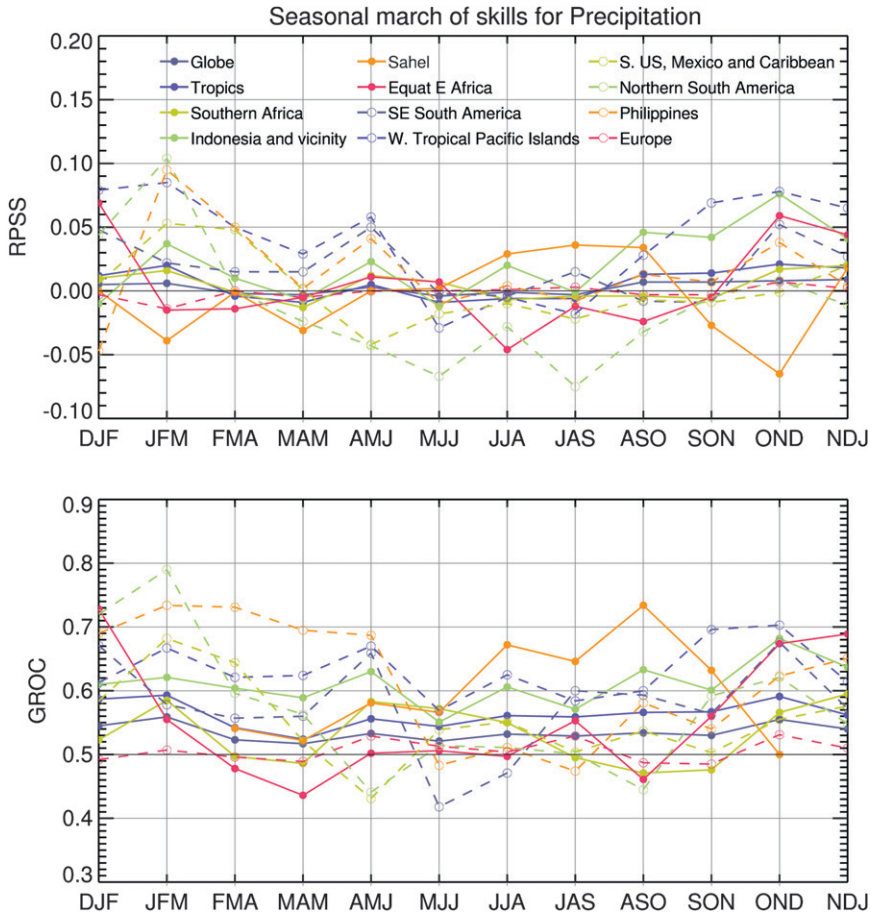

FIG. 7. Seasonal distribution of mean skill (1997–2008) for precipitation at 0.5-month lead time for selected regions, using (top) RPSS and (bottom) GROC. Rectangular boundaries: tropics (25°N–25°S); southern Africa [15°S southward (includes Madagascar)]; Indonesia and vicinity (20°N–10°S, 95°–150°E); Sahel (10°–20°N, 20°W–30°E); equatorial East Africa (10°N–10°S, 30°E eastward); southeast South America (25°–40°S, 60°W eastward); western tropical Pacific islands (23°N–25°S, 157°E–180°); southern United States, Mexico, and Caribbean (17°–32°N, 60°–120°W); northern South America (0°–12°N); Europe (35°–70°N, 55°E westward).

but not in GROC unless, additionally, the forecast warm tendency differs interannually in phase with the observations. Figure 7 (precipitation) and Fig. 9 (temperature) provide opposing examples of how attributes other than discrimination hurt or help RPSS, respectively, without affecting GROC. A summary comparison of skill results using the RPSS, LSS, and GROC measures is provided in Table 6 for each of the four forecast lead times for the globe, tropics, and extratropics.

d. Probabilistic reliability

Reliability plots for precipitation and temperature forecasts over the globe and in the tropics over the 11-yr forecast period, aggregated for all area-weighted grid points and seasons, are shown in Fig. 10 by tercile category. Precipitation reliability appears favorable, with very slight overconfidence for above- and below-normal precipitation. There is slight underforecasting of below-normal rainfall: the 11-yr verification period was slightly drier than the 30-yr climatological base period [e.g., in tropics, below- (above-) normal precipitation occurred in 36% (31%) of cases], while the mean forecast probabilities remained close to 33%. The frequency of issuance of given forecast probabilities (lower subpanels of Fig. 10) shows a majority of climatological probabilities forecasts both globally and for the tropics and non-climatology forecast probabilities deviating mainly within 10% of climatology. The near-unity slope of the reliability curve indicates that this lack of forecast sharpness is necessary, given the considerable forecast uncertainty consistent with the known generally limited skill levels. A summary of some diagnostic attributes of the reliability

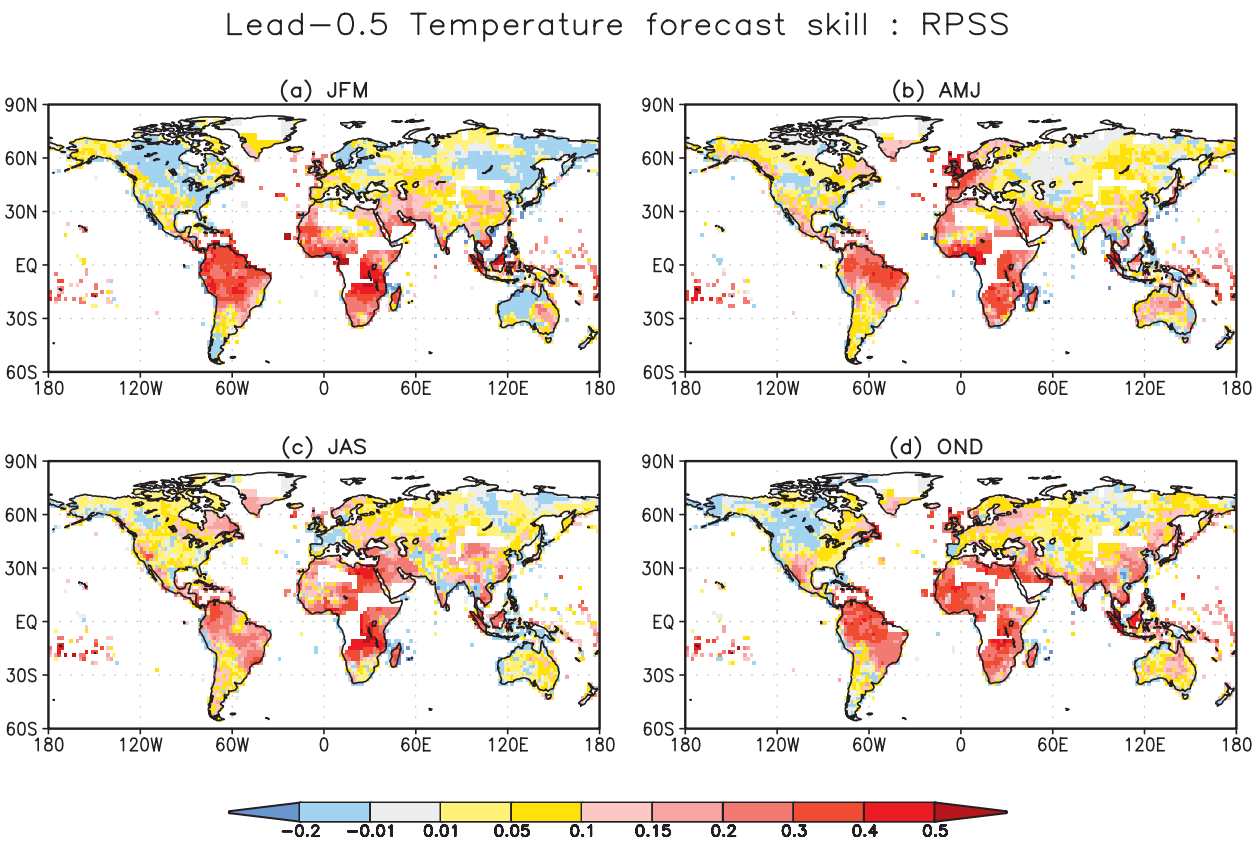

FIG. 8. Geographical distribution of mean RPSS for temperature for JFM, AMJ, JAS, and OND seasons at 0.5-month lead time. White areas lack sufficient data to calculate RPSS meaningfully.

analysis for the tropical precipitation forecasts is shown in Table 7a. Noted are reasonable slopes for the above- and below-normal categories and low sharpness (SD_f column).

For temperature (Fig. 10, bottom), the confidence component of the reliability is favorable, with reliability curve slopes near unity for above-normal temperature and slightly less for below normal for global and tropical domains. However, despite mean forecast probability shifts toward above-normal temperatures [mean probabilities issued for above (below) normal in tropics of 43% (23%)], above-normal temperatures were markedly underforecast [above (below) normal observed in 68% (10%) of cases]. This degree of imbalance in the observations with respect to the climatology reflects the large magnitude of low-frequency variability, including specifically a global warming signal.

The magnitude by which temperatures during the 11-yr study period averaged higher than those during the warmest of the 30-yr climatological base period used for IRI's forecasts (1971–2000) is illustrated in Fig. 11a, which shows the spatial distribution of the percentile of the 1998–2008 median temperature within the 1971–2000 climatologies, seasonally aggregated. Positive shifts are large: the 11-yr medians of 15% of the land grid points attain ≥95 percentile rank within the 1971–2000 observations, and the medians of 1.5% of the grid points are higher than *all 30 years* in the 1971–2000 period.[3] No grid points attain ≤5 percentile rank with respect to either 30-yr period, and most of the few grid points ranking below the median for 1971–2000 are near coastlines, restrained by the more slowly changing ocean temperatures. Throughout this period, during which the SST forecast models and AGCMs still used fixed, outdated greenhouse gas concentration settings, the IRI forecasts may have kept better pace with the warming trend if they had allowed an empirical tool known as optimal climate normals (OCN; Huang et al. 1996) to influence the forecasts.[4] Figure 11b shows the spatial distribution

[3] For the 1961–90 climatology used during the first few years of IRI forecasts, these figures increase to 30% and 8.6%, respectively.

[4] OCN forecasts the mean temperature anomaly observed over the most recent 10 years for the season and location in question and would forecast a pattern qualitatively similar to that of Fig. 10a, but season specific.

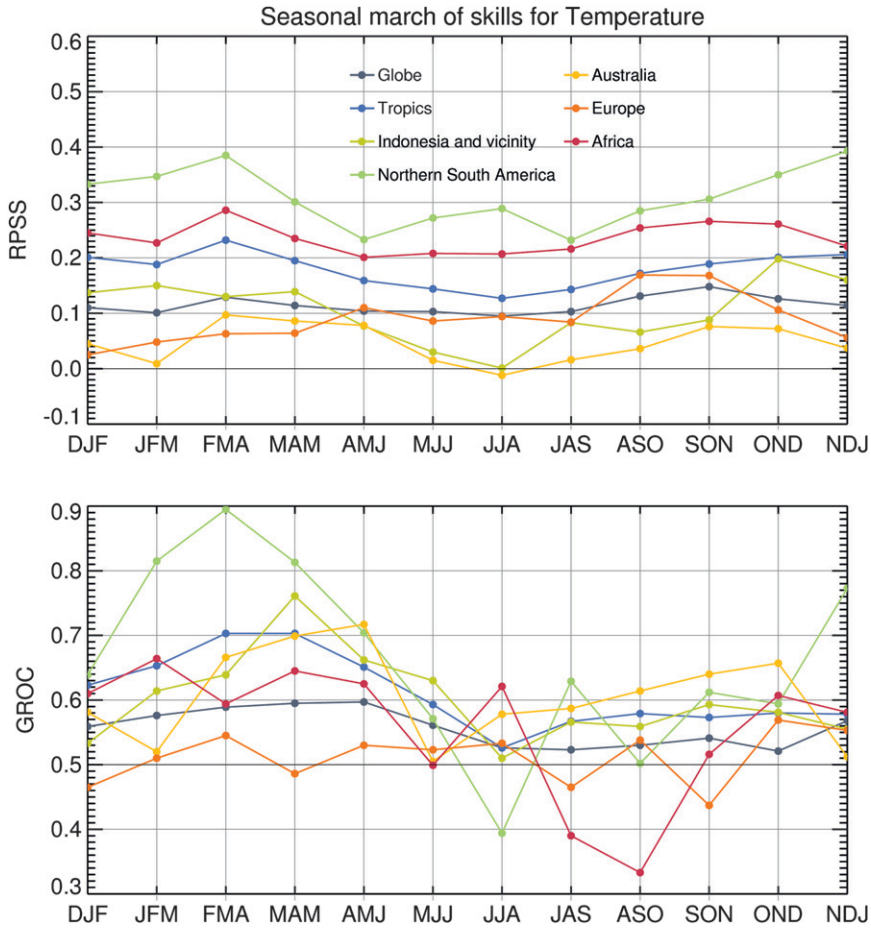

FIG. 9. Seasonal distribution of mean skill (1997–2008) for temperature at 0.5-month lead time for selected large regions, using (top) RPSS and (bottom) GROC. Rectangular boundaries for portions of continents are provided in caption of Fig. 7.

of the percentile of 1997–2008 median precipitation within the 1971–2000 climatology. The direction of shift from the climatology is geographically dependent for precipitation, with roughly equal areas trending drier as trending wetter.

Reliability plots for global temperature and precipitation (Fig. 10, left) show similar slopes, and slightly milder unconditional biases, compared with those for tropical forecasts. The global plots show lower forecast sharpness, consistent with known lower average signal-to-noise ratios (and expected predictive skill levels) in the extratropics (Shukla and Kinter 2006; Kumar et al. 2007; Peng et al. 2000).

A diagnostic evaluation of IRI's seasonal climate forecasts for the 1997–2000 period (Wilks and Godfrey 2002) found IRI's 0.5-month lead temperature forecasts somewhat overconfident, precipitation forecasts with appropriate confidence in the tropics but overconfidence in the extratropics, and substantial overforecasting of below-normal temperatures with a gross preponderance of above-normal observed temperature but only a slight mean tilt toward above-normal forecast temperature.[5] The negative RPSS values seen over a large region in extratropical latitudes (Figs. 5 and 6) suggest that even the weak shifts in precipitation probabilities have continued to exceed those warranted in the extratropics, and that nonclimatology forecasts should be uncommon in the extratropics. Mean tropical forecast probabilities for above-normal temperature were somewhat higher during the 11-yr forecast period than during 1998–2000 [partly in response to Wilks and Godfrey (2002) and because of forecasters' increasing confidence in forecasting a continuation of the global warming signal], while the relative frequency of observations in the above-normal category continued at the same high level (roughly

[5] A similarly strong cool bias was also noted in the climate forecasts of NOAA Climate Prediction Center during 1995–98 (Wilks 2000).

TABLE 6. RPSS, LSS, and GROC all-season skill results for IRI's precipitation and temperature forecasts for each of four lead times for land areas over the globe, tropics, and extratropics.

		Precipitation			Temperature		
Lead (months)	Measure	Globe	Tropics	Extratropics	Globe	Tropics	Extratropics
0.5*	RPSS	0.006	0.011	0.001	0.113	0.174	0.064
	LSS	0.003	0.006	0.000	0.054	0.087	0.028
	GROC	0.537	0.562	0.518	0.565	0.619	0.522
1.5	RPSS	0.004	0.006	0.002	0.087	0.133	0.050
	LSS	0.002	0.003	0.001	0.039	0.063	0.020
	GROC	0.528	0.547	0.513	0.553	0.606	0.511
2.5	RPSS	0.003	0.004	0.002	0.074	0.113	0.043
	LSS	0.001	0.002	0.000	0.032	0.052	0.016
	GROC	0.521	0.536	0.509	0.552	0.602	0.512
3.5	RPSS	0.003	0.006	0.001	0.059	0.094	0.031
	LSS	0.001	0.003	0.000	0.026	0.042	0.013
	GROC	0.522	0.538	0.510	0.543	0.586	0.509

* For comparison with the verification measure used for NOAA/Climate Prediction Center's seasonal forecasts for the United States (O'Lenic et al. 2008, Livezey and Timofeyeva 2008), Heidke skill scores for 0.5-month lead for precipitation for the globe, tropics, and extratropics are 0.037, 0.060, and 0.019, respectively, and for temperature they are 0.297, 0.408, and 0.208, respectively.

two-thirds of cases) over the 11-yr period as during 1998–2000. The result was a slightly less severe, but still very substantial, cool bias.

e. Separation of interannual and low-frequency skill

The performance of probabilistic forecasts is more fully described by verification measures aimed at different attributes than by a single measure (Wilks 2006). For example, RPSS and LSS can be negative because of imperfect calibration even when the forecasts have potential information value (Hsu and Murphy 1986; Mason 2004), while GROC, being virtually insensitive to calibration problems (e.g., mean or conditional forecast biases), may show positive results for the same set of forecasts. Among the RPSS, LSS, GROC scores, and reliability diagrams, multidimensional diagnostics are formed for the forecasts.

A comparison among the geographical patterns of RPSS, LSS, and GROC is shown for IRI's forecasts of precipitation and temperature for the JFM season in Figs. 12 and 13, respectively. Comparisons for other seasons (not shown) are similar. RPSS and LSS have very similar patterns, including locations having zero skill, LSS averaging one-half to one-third of RPSS in magnitude. That RPSS is affected by probabilities assigned to categories that do not verify, while LSS is not, is probably not a significant factor in the score differences for IRI's forecasts, which never have grossly non-Gaussian (e.g., bimodal) probabilities that would enable probabilities given to nonverifying categories to be important. The global spatial correlation between RPSS and LSS is 0.9 or greater for all seasons, while that between either of them and GROC is approximately 0.6. Thus, at least for IRI forecast skill, RPSS and LSS appear largely redundant, and either of them could be used alone without material loss of information. For precipitation (Fig. 12), GROC skill shows skill patterns roughly similar to those of RPSS and LSS, but with somewhat less area of negative (<0.5) skill.[6] This suggests that the proportion of correct discrimination among the varying probabilities forecast for the tercile categories is favorable in JFM in the relatively high skill regions (e.g., Philippines—east Australia, Pacific islands). Because forecast uncertainty is considerable (the most likely category often having only 0.40–0.50 probability, as warranted for good reliability), RPSS and LSS have weak magnitudes in these skillful regions. Additionally, the mild bias of over- (under-) forecasting above (below) normal further decreases RPSS and LSS but not GROC (Fig. 10 and Table 7a); these probabilistic features may cause RPSS and LSS to be negative.

A different picture is presented for temperature (Fig. 13). The patterns themselves, while roughly similar, differ in that there is a greater area of negative skill for GROC than for RPSS and LSS. This is caused by the

[6] The larger spatial variation seen in GROC within its range of 0 to 1 than those seen in RPSS and LSS within their ranges exists, in part, because GROC measures mainly just one attribute of performance (discrimination), while RPSS and LSS measure performance in discrimination and in other attributes. Excellent (or very poor) performance is less likely to occur in net over several attributes than in just one.

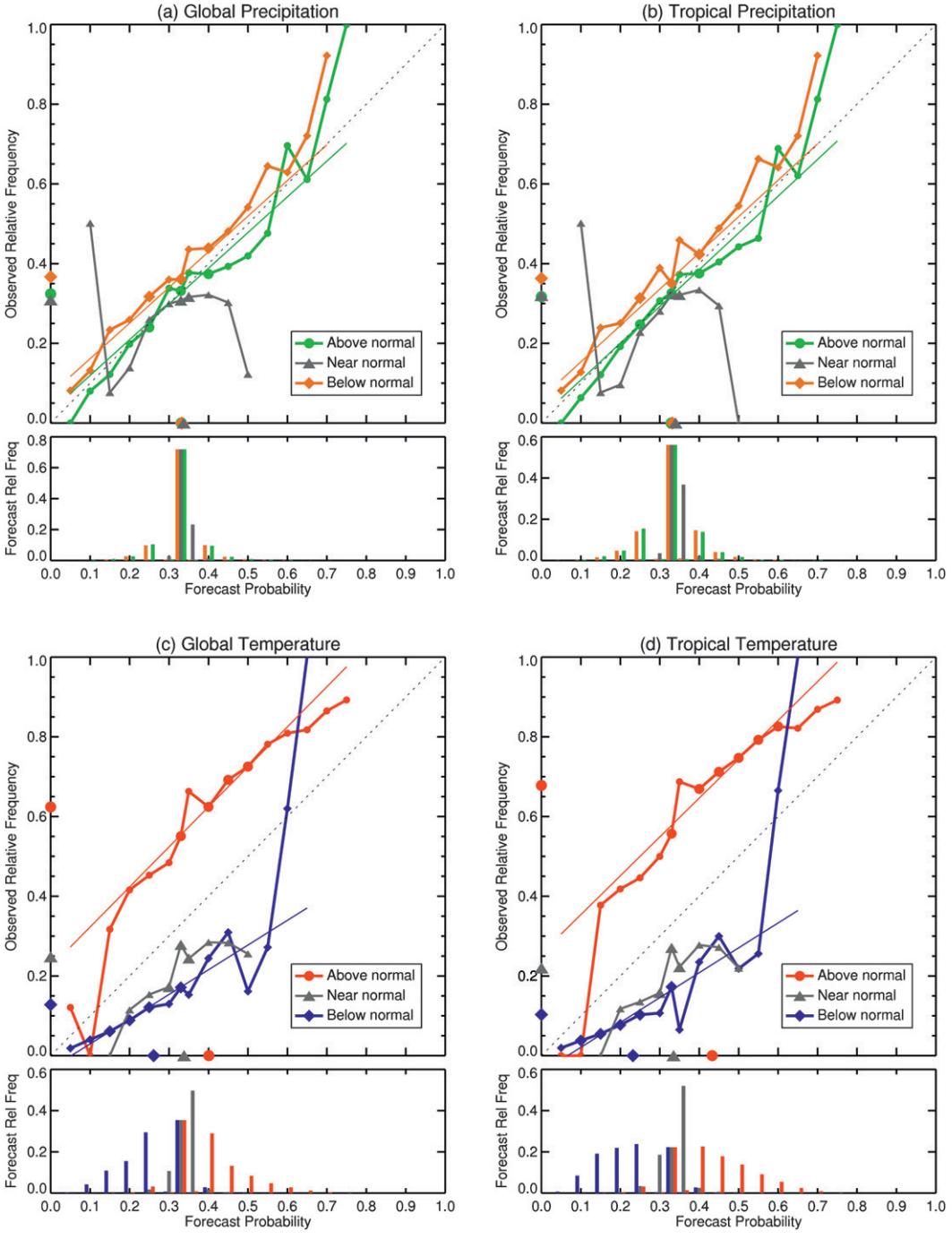

FIG. 10. Reliability plot for (a) global and (b) tropical precipitation and (c) global and (d) tropical temperature for 0.5-month lead forecasts for all seasons. For precipitation, the green curve pertains to forecast probabilities for above-normal precipitation, the orange curve pertains to forecast probabilities for below-normal precipitation, and the gray curve pertains to forecast probabilities for near-normal precipitation. For temperature, the red curve denotes above-normal temperature and blue below-normal temperature. For above and below normal, least squares regression lines are shown, weighted by the sample sizes represented by each point. Points representing probability intervals that are forecast at least 5% of the time are drawn using larger symbols than other points. The diagonal $y = x$ line represents perfect reliability. The colored marks on the axes show the overall means of the forecast probabilities or observed relative frequencies. The lower part of each panel shows the frequency with which each interval of probability was forecast, where interval widths are 0.05 (e.g., 0.175–0.225 is labeled as 0.20), except that the climatological (0.333) probability is also explicitly shown.

TABLE 7. Elements of diagnostic evaluation of skill and reliability of IRI forecasts over the tropics during 1997–2008 for (a) issued forecasts of precipitation, (b) multimodel ensemble forecasts of precipitation, (c) issued forecasts of temperature, and (d) multimodel ensemble forecasts of temperature. Forecast probabilities issued for each tercile-based category are diagnosed in each part, and overall verification scores are shown in each panel heading. SD denotes std dev, res denotes resolution, rel reliability, BS Brier score, BSS Brier skill score; subscripts *f* and *c* denote forecast and climatology reference forecast, respectively.

(a) Tropical precipitation —issued forecasts (RPSS 0.011, LSS 0.006, GROC 0.562)

Category	Obs rel frequency	Avg forecast	Bias	Slope	Intercept	SD_f	Res	Rel_f	Rel_c	BS_f	BS_c	BSS	ROC
Wet	0.317	0.329	0.012	0.922	0.016	0.070	0.0044	0.0003	0.0003	0.218	0.222	0.020	0.566
Near normal	0.320	0.340	0.020	0.424	0.177	0.021	0.0003	0.0007	0.0002	0.223	0.222	−0.001	0.505
Dry	0.363	0.332	−0.031	0.912	0.062	0.071	0.0046	0.0015	0.0009	0.219	0.223	0.018	0.572

(b) Tropical precipitation ——multimodel ensemble forecasts (RPSS −0.019, LSS .−004, GROC 0.555)

Category	Obs rel frequency*	Avg forecast	Bias	Slope	Intercept	SD_f	Res	Rel_f	Rel_c	BS_f	BS_c	BSS	ROC
Wet	0.313	0.339	0.026	0.451	0.149	0.110	0.0029	0.0042	0.0004	0.224	0.223	−0.004	0.563
Near normal	0.316	0.339	0.023	0.133	0.279	0.066	0.0002	0.0037	0.0003	0.226	0.223	−0.014	0.508
Dry	0.371	0.324	−0.047	0.443	0.226	0.107	0.0024	0.0054	0.0014	0.225	0.224	−0.007	0.550

(c) Tropical temperature—issued forecasts (RPSS 0.174, LSS 0.087, GROC 0.619)

Category	Obs rel frequency	Avg forecast	Bias	Slope	Intercept	SD_f	Res	Rel_f	Rel_c	BS_f	BS_c	BSS	ROC
Warm	0.678	0.433	−0.245	0.974	0.256	0.097	0.0097	0.0605	0.1188	0.273	0.341	0.199	0.624
Near normal	0.220	0.335	0.115	0.875	−0.073	0.030	0.0017	0.0143	0.0128	0.235	0.235	0.001	0.529
Cool	0.103	0.232	0.129	0.624	−0.042	0.081	0.0027	0.0178	0.0535	0.237	0.276	0.139	0.650

(d) Tropical temperature—multimodel ensemble forecasts (RPSS 0.188, LSS 0.102, GROC 0.610)

Category	Obs rel frequency *	Avg forecast	Bias	Slope	Intercept	SD_f	Res	Rel_f	Rel_c	BS_f	BS_c	BSS	ROC
Warm	0.689	0.481	−0.208	0.514	0.454	0.174	0.0091	0.0640	0.1265	0.277	0.349	0.205	0.611
Near normal	0.213	0.313	0.100	0.291	0.122	0.109	0.0012	0.0193	0.0145	0.240	0.237	−0.015	0.547
Cool	0.098	0.206	0.108	0.331	0.030	0.124	0.0017	0.0220	0.0554	0.243	0.278	0.126	0.609

* Observed relative frequency for multimodel ensemble forecasts differs slightly from that of issued forecasts because of slightly differing sets of grid squares forecast between the two forecast sets.

recognition, albeit too weak, of the dominance of the above-normal category in the forecasts that is rewarded in RPSS and LSS but not in GROC, where only correct discrimination among the sampled cases is credited. Figure 13 shows that, while discrimination among the temperature forecasts within the 11-yr forecast sample was better for temperature than for precipitation in the tropics, it was generally low for both temperature and precipitation outside of the tropics. The level of discrimination for the outer categories for tropical temperature is indicated by ROC scores in the low to middle 0.60s (Table 7), compared with the upper 0.50s for tropical precipitation.

To summarize, the GROC helps to distinguish forecast skill related solely to discrimination of mainly interannual variability within the 11-yr forecast period, as opposed to such discrimination combined with skill in correctly predicting overall 11-yr mean probability shift with respect to the 30-yr climatological base period(s), as for example, that associated with climate change. GROC shows minor differences from RPSS and LSS (allowing for scaling differences) for precipitation (Fig. 12), being more favorable because it is not penalized for the small wet bias. For temperature (Fig. 13), GROC shows smaller areas of positive skill than RPSS and LSS because the climate shift in temperature was large enough that even partial recognition of it in IRI's probability forecasts (Fig. 10 and Table 7) was credited in RPSS and LSS but not GROC. However, even for discrimination alone, performance is seen to be stronger for temperature than for precipitation in the tropics.

f. Skill of objective multimodel predictions; comparison with issued forecasts

A comparison of the skill of IRI's issued forecasts with that of the objective multimodel ensemble forecasts indicates the value of the human modification to the raw model output. The objective output comes from several AGCMs, each roughly calibrated to its mean and terciles, forced by multiple SST scenarios, and weighted using two multimodel ensemble algorithms. Ideally, the objective probabilistic model output should be capable of being the final forecast product, but expert judgment has further influenced the issued forecasts. As discussed earlier, subjective modifications include a general weakening of probability anomalies, more specific weakening

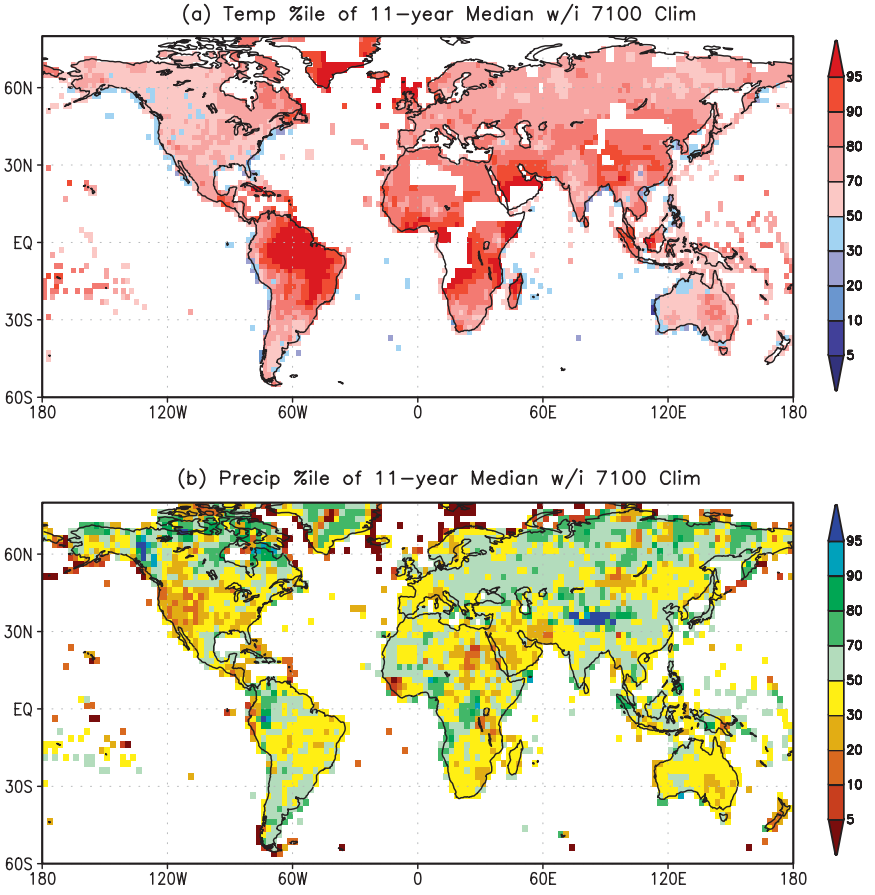

FIG. 11. Geographical distribution of decadal (or lower frequency) climate change: (a) percentile of 11-yr median temperature in terms of the 1971–2000 climatology; (b) percentile of 11-yr median precipitation in terms of the 1971–2000 climatology. White areas in (a) lack sufficient data to calculate the percentile meaningfully.

of excessively sharp forecasts, spatial smoothing, spatial MOS corrections for selected regions/seasons for precipitation, shifting of temperature probabilities toward "above normal," and adjustments toward forecasts issued by other producing centers. Weakening of forecast probability anomalies is done because the model weighting scheme in the multimodel combination is based on historical skills using observed rather than predicted SST.[7] Probabilistic shifts toward above normal for temperature are done because the models do not fully capture global warming, even with warmer SSTs forcing them, because of constant and outdated prescribed model greenhouse gas concentrations.

Figure 14 shows the geographical distribution of RPSS for the multimodel precipitation and temperature predictions at 0.5-month lead for JFM and JAS. While a comparison of the precipitation skills to those of the actually issued forecasts (Figs. 6a,c and 3 and Tables 7a,b) indicates similar skills for both variables, differences are discernible. For both seasons, spatially noisier skill patterns are seen for the multimodel than for the issued precipitation forecasts. Global and tropical mean skills for JFM and JAS precipitation are very slightly higher for the issued than for the multimodel forecasts, and inspection of the two RPSS fields suggests that this may be largely due to the smoothing and weakening of predicted deviations from climatological probabilities. The level of spatial noise in the multimodel forecasts is greater for precipitation than for temperature (Gong et al. 2003), requiring more forecaster smoothing to optimize skill. The same comparison using the GROC score (not shown) leads to a qualitatively similar conclusion, except with even smaller skill differences between the two forecast

[7] Although hindcasts using SST empirically predicted using constructed analog (Van den Dool 1994, 1997) and persisted from observed SST (Li et al. 2008) are available for the AGCMs run at IRI (ECHAM and CCM), they have not been generated for the AGCMs run at partner institutions.

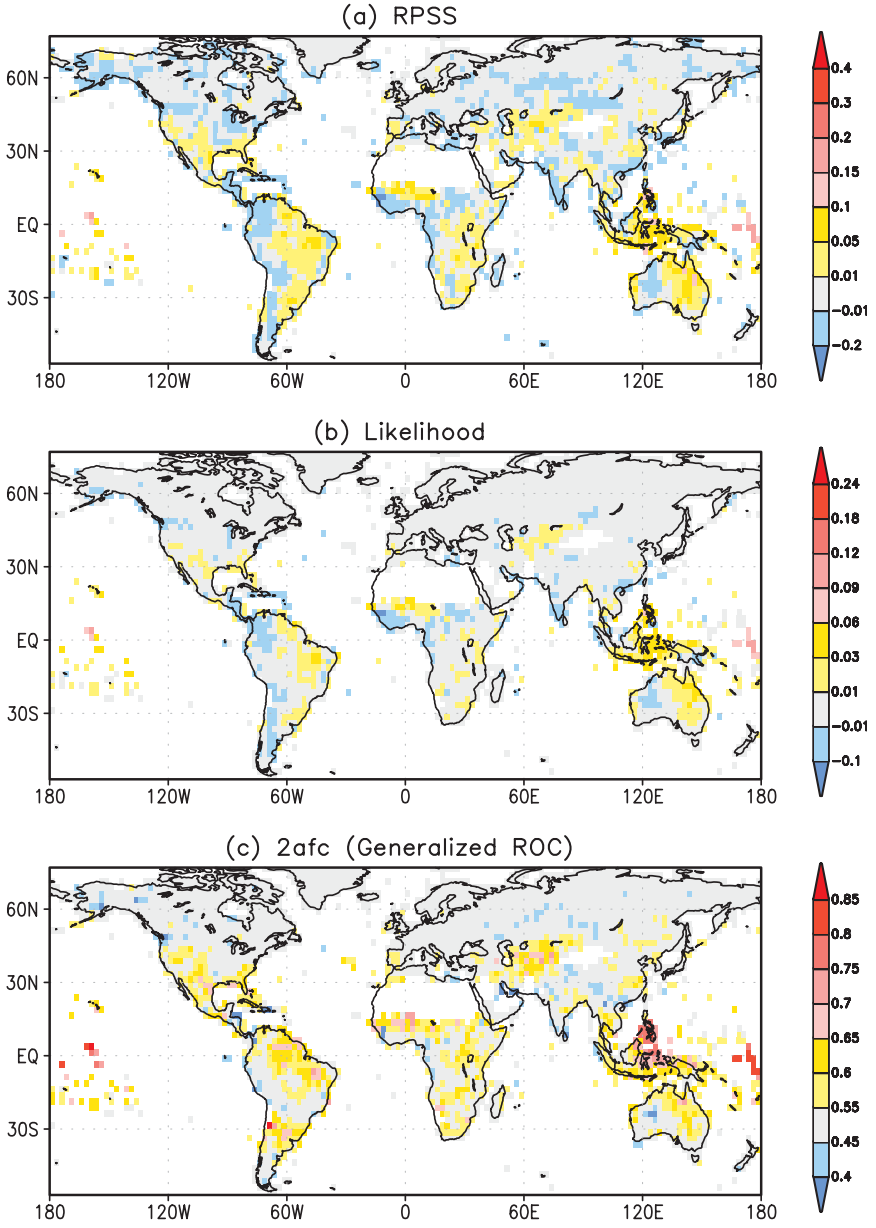

FIG. 12. Geographical distribution of all-season precipitation skill verified using (a) RPSS, (b) LSS, and (c) GROC. White areas lack sufficient data to calculate RPSS meaningfully.

sets, likely because of approximately equal levels of basic discrimination in both forecast versions, but better calibration in issued than multimodel forecasts.

With relatively small trend components in precipitation, interpretation of results is related mainly to interannual variability. The reliability plot for all-season tropical precipitation multimodel ensemble predictions (Fig. 15, comparable to Fig. 10 for issued forecasts) shows somewhat shallower slopes, indicative of greater overconfidence in the multimodel ensemble predictions. Tables 7a,b provide attributes of reliability and skill for the 0.5-month lead tropical precipitation multimodel forecasts and issued all-season forecasts. The issued probabilities are more conservative and have very slightly higher skills by most of the verification measures. The mean squared departures of the reliability curve from the ideal 45° line (“reliability” column in Table 7) are greater, and exceed those of the climatology forecast reference by greater amounts, for multimodel ensemble forecasts than for issued forecasts for all three categories. A somewhat higher resolution is also seen in the issued than the multimodel forecasts and is closely related

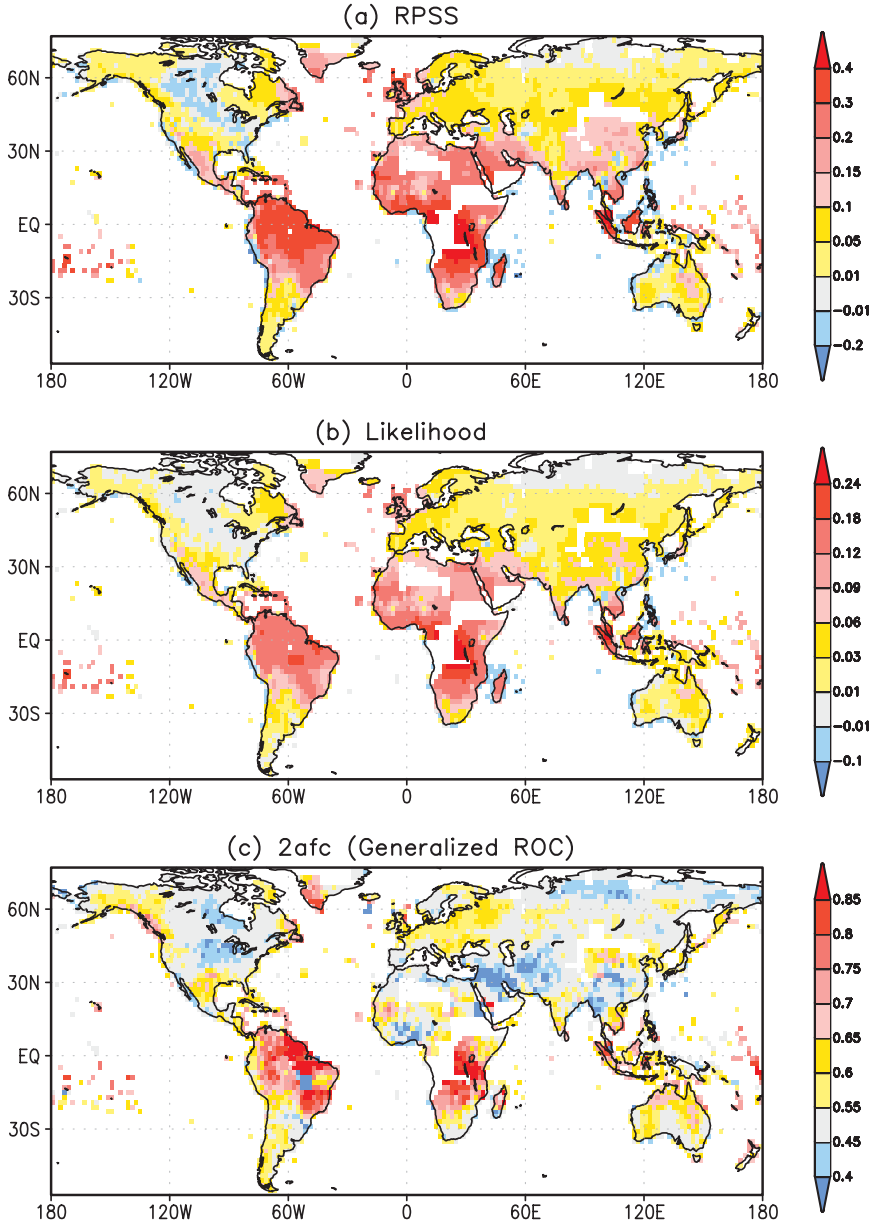

FIG. 13. Geographical distribution of all-season temperature skill verified using (a) RPSS, (b) LSS, and (c) GROC. White areas lack sufficient data to calculate RPSS meaningfully.

to small increases in discrimination (indicated by greater GROC and ROC by individual category), which is believed to be due to the forecasters' additional spatial smoothing and modifications resulting from the selected regional spatial MOS corrections.

The right side of Fig. 14 shows the geographical distribution of RPSS for multimodel temperature predictions at 0.5-month lead for JFM and JAS. Comparison with skill for the corresponding issued forecasts (Figs. 8a,c and 3) shows for both seasons skill patterns of roughly similar mean level, but spatially noisier, in the multimodel forecasts. However, global and tropical mean RPSS are just slightly higher in the multimodel than the issued forecasts (Figs. 3 and 14 and Tables 7c,d), mainly because of higher scores in those tropical regions where skill is highest in both forecast versions. In regions of low extratropical skill, issued forecasts have milder negative RPSS than multimodel forecasts. A comparison between objective and finally issued forecasts using GROC (not shown) indicates no edge in performance of the multimodel ensemble forecasts over the issued forecasts. A likely explanation is that the forecaster modifications change the calibration of

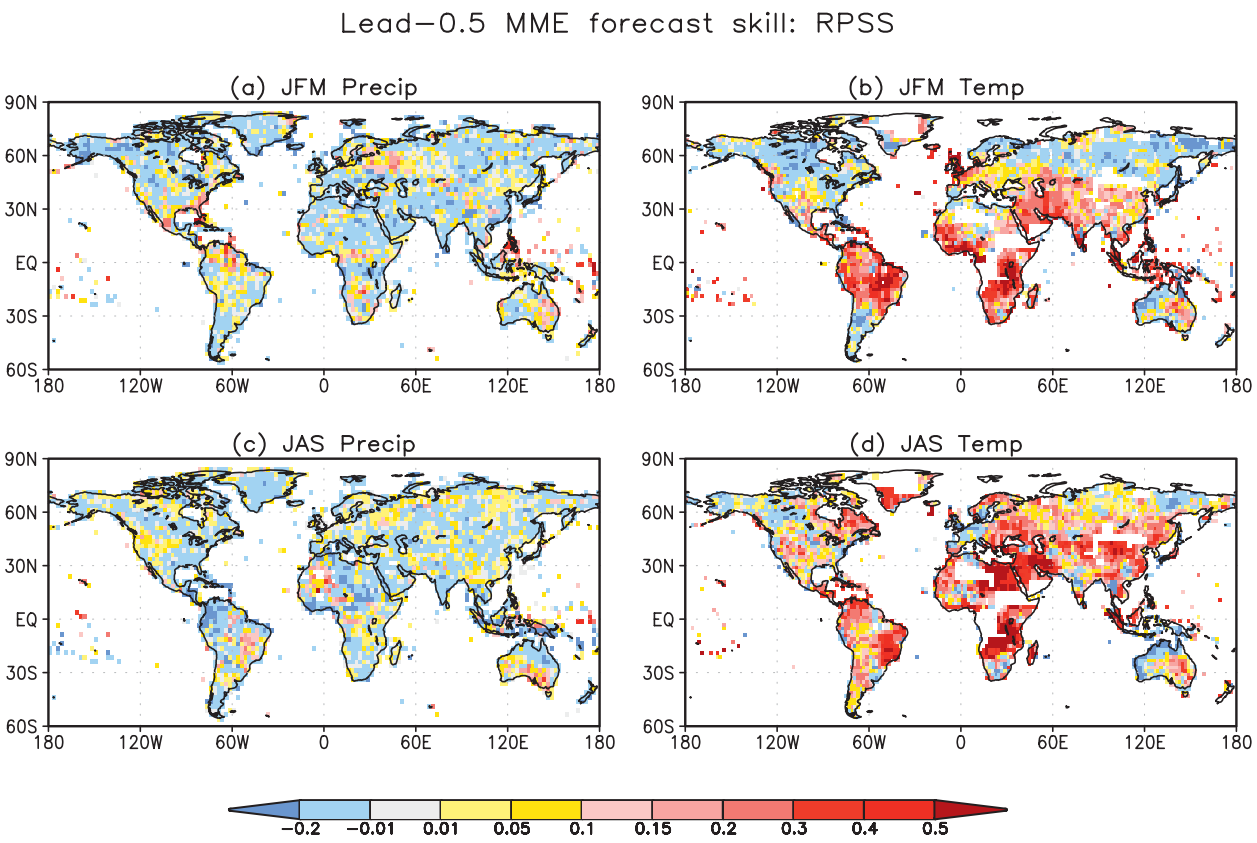

FIG. 14. Geographical distribution of RPSS for unmodified multimodel predictions at 0.5-month lead for JFM for (a) precipitation and (b) temperature and for JAS for (c) precipitation and (d) temperature. White areas lack sufficient data to calculate RPSS meaningfully.

the forecasts (Fig. 15 versus Fig. 10; Tables 7c,d) but have little effect on the basic discrimination present in the multimodel ensemble forecasts. However, the combination of decreasing the confidence and weakly adjusting for the underforecasting of above-normal temperature, while leaving GROC unchanged as expected, slightly reduced RPSS. Although the forecasters' adjustments for the cool bias increased RPSS, their reduction of perceived overconfidence had a larger negative impact on RPSS by weakening the highest probabilities for above normal. While RPSS and LSS are slightly lower for the issued than multimodel ensemble temperature forecasts, the GROC, the ROC for individual categories, and the resolution components (Tables 7c,d) suggest slight improvements in discrimination in the issued forecasts and improved slopes in the reliability curves.

In summary, overconfidence in the multimodel forecasts was corrected in the issued forecasts for both precipitation and temperature, but the objective forecasts better captured the strong warming trend than the issued forecasts. The issued forecasts featured equal to slightly improved resolution/discrimination compared with the model output, but the damping of above-normal probabilities to correct for overconfidence contributed to its underforecasting.

5. Summary

The IRI has issued seasonal probabilistic forecasts of near-global temperature and precipitation for 11 years since late 1997, using mainly a two-tiered, dynamically based prediction system where a set of SST prediction scenarios is made, which then serve as prescribed lower boundary conditions for integrations of ensembles from a set of AGCMs. Forecasts have been issued monthly, for four upcoming running 3-month periods, for most of the IRI's history. Seven AGCMs have been used since 2004, whereby forecast ensembles numbering well over 100 members are postprocessed and merged into final probability forecasts.

The skill of the forecasts ranges from near zero to moderate, depending on season and location. Skills for temperature average higher, are less seasonally and regionally dependent, and decay more slowly with lead time than skills for precipitation. Temperature skills benefit from correct forecasts of continuation of a strong tendency

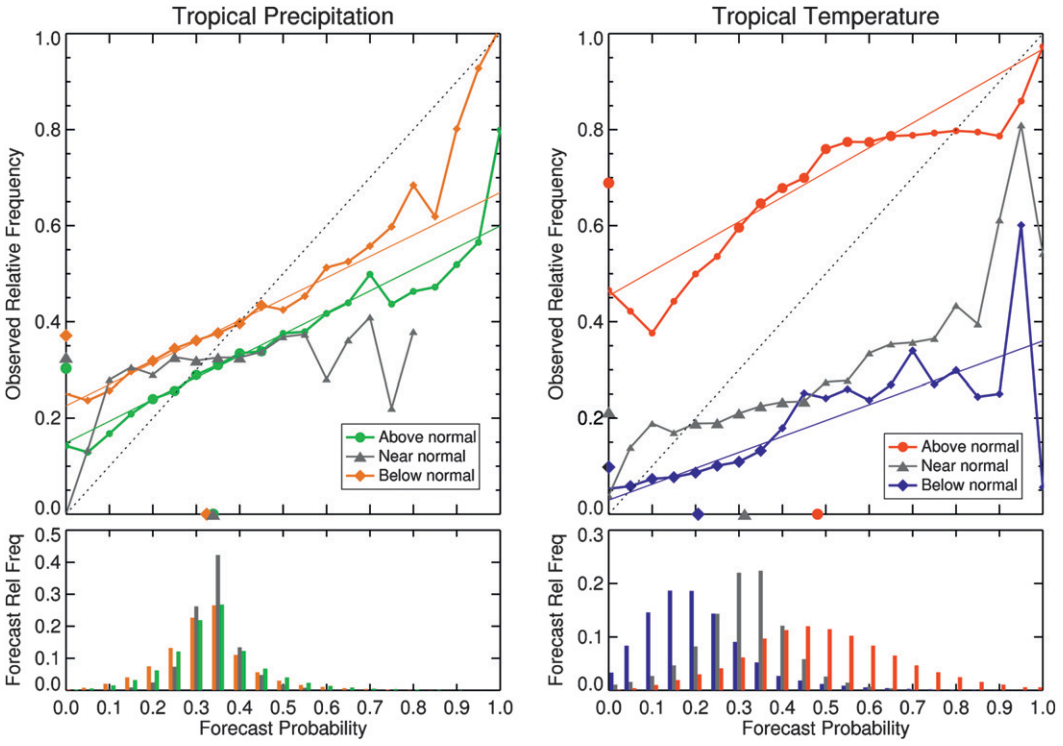

FIG. 15. Reliability plot for tropical (left) precipitation and (right) temperature for 0.5-month lead unmodified multimodel predictions for all seasons. Green (red) denotes above normal and orange (blue) denotes below normal for precipitation (temperature), and gray denotes near normal. Details are as described in caption of Fig. 10. Forecasts for climatological probabilities (0.333) are not shown explicitly.

for above-average temperatures (relative to a completed 30-yr base period) associated with global warming. Although ENSO remains a source of temperature forecast skill, warming trends have rivaled ENSO effects as a skill source during the forecast period. Skills for precipitation, by contrast, do not benefit appreciably from a trend component because precipitation trends are weaker and vary in direction depending on season and location. Hence, precipitation skill is based mainly on correctly discriminated effects of interannual fluctuations involving ENSO and SST anomalies outside the tropical Pacific.

Forecast skills are higher in the tropics than extratropics for both temperature and precipitation. This is consistent with the higher signal-to-noise ratios at low latitudes documented for troposphere geopotential height (e.g., Shukla and Kinter 2006; Kumar et al. 2007) and in associated surface climate (e.g., Rowell 1998; Peng et al. 2000). While the spatial pattern of temperature forecast skill shows a weak annual cycle, that of precipitation is more strongly seasonally dependent, roughly following both the annual cycle of low-latitude monsoon rainfall and teleconnections to large-scale tropical SST anomalies—particularly ENSO. At midlatitudes, positive precipitation skill, while not prevalent, is found in regions and seasons having successfully modeled ENSO and non-ENSO tropical SST teleconnections. The skill results found here are consistent with skill evaluations by other forecast-producing centers and with theoretical predictability studies. Skill levels in specific seasons and locations could benefit users who understand the probabilistic aspects of seasonal climate forecasts sufficiently for prudent decision making for their application.

Over a period as brief as 11 yr, the variability in the amplitude of ENSO extremes is likely to govern forecast skill more strongly than incremental improvements in models or forecast methodology. Hence, Fig. 3 indicates highest forecast skills during the 1997/98 El Niño at the beginning of the period, despite the fact that the simplest SST prediction scheme and the fewest AGCMs were used at that time.

Skills of the objective multimodel probability forecasts, used as the primary basis for the final issued forecasts, are comparable to those of the final forecasts, but they are somewhat overconfident. This is believed to be due in part to the development of the multimodel superensembling process using individual AGCM skills when the AGCMs are forced by observed rather than

predicted SST. Thus, while the relative weighting among the models may be well estimated, their collective weighting, and resultant departures from climatological forecast probabilities, are overestimated.

The verification diagnostics challenge the suitability of using completed 30-yr periods to define the current temperature climatology from which to form anomalies or quantile-based category boundaries, given the strong nonstationarity. Reasons to consider alternative climatological reference frames include the severely shifted categorical frequencies of current observations, forecasts reflecting a mixture of time scales that may be confusing to stakeholders, and the greater challenge in conducting meaningful verification. Observational alternatives to estimation of the current year's temperature climate might include use of an annually updated OCN-based climatology (Huang et al. 1996) or, at higher risk, a linear trend fit of the observations in recent decades (e.g., Livezey et al. 2007); a dynamical approach might consist of a deemed optimal superensemble of regionally specific Intergovernmental Panel on Climate Change (IPCC) model forecasts (e.g., Tebaldi et al. 2005; Greene et al. 2006; Furrer et al. 2007; Christensen et al. 2007; Gleckler et al. 2008) averaged over a period centered on the current year. The above options all carry uncertainties beyond those of a stationary climate, as they contain predictive components.

The aspects of IRI's prediction system in greatest need of improvement or further development are 1) postprocessing: use of systematic spatial MOS corrections for individual AGCMs, specific to the season and lead time, before superensembling; 2) incorporation of time-varying greenhouse gas settings in SST forecast models and in AGCMs; and 3) movement toward a partially or totally single-tiered prediction system. Implementation of 1) occurred in late 2009, and progress on 3) is under way.

It is difficult to compare the operational predictive skill of IRI's forecast system with that of other systems such as a single-tiered dynamical system [e.g., Palmer et al. 2004 (and references therein); Graham et al. 2005; Saha et al. 2006; Kug et al. 2008; Wang et al. 2008] or a purely empirical system (Van den Dool 2007). Improvement in ENSO prediction has obvious value toward improvement of climate prediction, and the potential predictability of ENSO is an open question but believed not fully realized (Chen and Cane 2008). Expansion of available data from the Argo (Schmid et al. 2007), Prediction and Research Moored Array in the Tropical Atlantic (PIRATA) (Bourlès et al. 2008), and Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction (RAMA) (McPhaden et al. 2009) systems is expected to result in more fully realizable predictive skill for SST in tropical oceans outside of the Pacific. Improved modeling of the ocean–atmosphere system, through better representation of physical processes, should increase skill toward the theoretical limit and reduce the need for postprocessing and forecaster intervention.

Acknowledgments. This work was funded by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA07GP0123 and NA050AR4311004). The views expressed are those of the authors and do not necessarily reflect the views of NOAA or its subagencies. The authors appreciate the thoughtful and constructive comments of three anonymous reviewers. The monthly forecast AGCM integrations done by partner institutions (NASA GSFC, COLA, Queensland Climate Change Centre of Excellence, GFDL, Scripps/ECPC) have been invaluable contributions to IRI's forecasts. Acknowledged are scientific contributions by Nicholas Graham for the original system design, Stephen Zebiak, Balaji Rajagopalan, Upmanu Lall, and Andrew Robertson for multimodel superensembling, and Michael Tippett for targeted spatial AGCM MOS corrections. Competent production support was provided by Mary Tyree, Martin Olivera, John del Corral, Jack Ritchie, Sara Barone, and Bin Li.

REFERENCES

Aldrich, J., 1997: R. A. Fisher and the making of maximum likelihood 1912–1922. *Stat. Sci.,* **12,** 162–176.

Alexander, M. A., I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott, 2002: The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans. *J. Climate,* **15,** 2205–2231.

Bacmeister, J., P. J. Pegion, S. D. Schubert, and M. J. Suarez, 2000: Atlas of seasonal means simulated by the NSIPP 1 atmospheric GCM. NASA/TM-2000-104505, Vol. 17, 194 pp.

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.,* **115,** 1825–1850.

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures: Refinement of the Heidke score. *Wea. Forecasting,* **7,** 699–709.

——, M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.–Ocean,* **35,** 367–383.

——, 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.

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.

Bourlès, B., and Coauthors, 2008: The PIRATA program: History, accomplishments, and future directions. *Bull. Amer. Meteor. Soc.,* **89,** 1111–1125.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Brocker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: The importance of being proper. *Wea. Forecasting,* **22,** 382–388.

Chang, P., and Coauthors, 2006: Climate fluctuations of tropical coupled systems—The role of ocean dynamics. *J. Climate,* **19,** 5122–5174.

Chen, D., and M. A. Cane, 2008: El Niño prediction and predictability. *J. Comput. Phys.,* **227,** 3625–3640.

——, ——, A. Kaplan, S. E. Zebiak, and D. Huang, 2004: Predictability of El Niño over the past 148 years. *Nature,* **428,** 733–735.

Christensen, J. H., and Coauthors, 2007: Regional climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 847–940.

Deutsches Klimarechenzentrum, 1992: The ECHAM-3 atmospheric general circulation model. Tech. Rep. 6, 189 pp. [Available from the Modellbetruungsgruppe, Deutsches desstr. 55, Hamburg D-20146, Germany.]

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.

Enfield, D. B., A. Mestas-Nuñez, D. A. Mayer, and S. L. Cid, 1999: How ubiquitous is the dipole relationship in the tropical Atlantic sea surface temperatures? *J. Geophys. Res.,* **104** (C4), 7841–7848.

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

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka, and G. A. Meehl, 2007: Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.,* **34,** L06711, doi:10.1029/2006GL027754.

GFDL Global Atmospheric Model Development Group, 2004: The new GFDL global atmosphere and land model (AM2-LM2): Evaluation with prescribed SST simulations. *J. Climate,* **17,** 4641–4673.

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.,* **25,** 23–31.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** D06104, doi:10.1029/2007JD008972.

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** (D16), 19 099–19 116.

——, and M. Dilley, 2005: El Niño: Catastrophe or opportunity. *J. Climate,* **18,** 651–665.

——, S. J. Mason, 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.

——, 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.

Graham, R. J., M. Gordon, P. J. McLean, S. Ineson, M. R. Huddleston, M. K. Davey, A. Brookshaw, and R. T. H. Barnes, 2005: A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. *Tellus,* **57A,** 320–339.

Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Climate,* **19,** 4326–4343.

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

Hagedorn, R., and L. A. Smith, 2008: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.,* **16,** 143–155, doi:10.1002/met.92.

Halpert, M. S., and C. F. Ropelewski, 1992: Surface temperature patterns associated with the Southern Oscillation. *J. Climate,* **5,** 577–593.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometric framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

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.

Hurrell, J. W., J. J. Hack, B. A. Boville, D. L. Williamson, and J. T. Kiehl, 1998: The dynamical simulation of the NCAR Community Climate Model version 3 (CCM3). *J. Climate,* **11,** 1207–1236.

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

Kanamitsu, M., and K. C. Mo, 2003: Dynamical effect of land surface processes on summer precipitation over the southwestern United States. *J. Climate,* **16,** 496–509.

——, and Coauthors, 2002: NCEP dynamical season forecast system 2000. *Bull. Amer. Meteor. Soc.,* **83,** 1019–1037.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model. *J. Climate,* **11,** 1131–1149.

Kug, J. S., I. S. Kang, and D. H. Choi, 2008: Seasonal climate predictability with tier-one and tier-two prediction systems. *Climate Dyn.,* **31,** 403–416.

Kumar, A., and M. P. Hoerling, 2003: The nature and causes for the delayed atmospheric response to El Niño. *J. Climate,* **16,** 1391–1403.

——, ——, M. Ji, A. Leetmaa, and P. Sardeshmukh, 1996: Assessing a GCM's suitability for making seasonal forecasts. *J. Climate,* **9,** 115–129.

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

——, B. Jha, Q. Zhang, and L. Bounoua, 2007: A new methodology for estimating the unpredictable component of seasonal atmospheric variability. *J. Climate,* **20,** 3888–3901.

Kushnir, Y., W. A. Anderson, P. Chang, and A. W. Robertson, 2006: The physical basis for predicting Atlantic sector seasonal-to-interannual climate variability. *J. Climate,* **19,** 5949–5970.

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

Lau, N.-C., and M. J. Nath, 1996: The role of atmospheric bridge in linking tropical Pacific ENSO events to extratropical SST anomalies. *J. Climate,* **9,** 2036–2057.

Li, S., L. Goddard, and D. G. DeWitt, 2008: Predictive skill of AGCM seasonal climate forecasts subject to different SST prediction methodologies. *J. Climate,* **21,** 2169–2186.

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.

Livezey, R. E., and W.-Y. Chen, 1983: Field significance and its determination by Monte-Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

——, 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.

——, 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.

——, K. Y. Vinnikov, M. M. Timofeyeva, R. Tinker, and H. M. Van den Dool, 2007: Estimation and extrapolation of climate normals and climatic trends. *J. Appl. Meteor. Climatol.,* **46,** 1759–1776.

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

Mason, S. J., 2004: On using ''climatology'' as a reference strategy in the Brier and ranked probability skill scores *Mon. Wea. Rev.,* **132,** 1891–1895.

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

——, and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate,* **15,** 8–29.

——, 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.

McPhaden, M. J., and Coauthors, 2009: RAMA: The Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction. *Bull. Amer. Meteor. Soc.,* **90,** 459–480.

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.

——, and E. S. Epstein, 1967: A note on probability forecasts and ''hedging.'' *J. Appl. Meteor.,* **6,** 1002–1004.

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.

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.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Pegion, P. J., S. D. Schubert, and M. J. Suarez, 2000: An assessment of the predictability of northern winter seasonal means with the NSIPP 1 AGCM. NASA/TM-2000-104505, Vol. 18, 110 pp.

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 the 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, 2004: Statistical prediction of sea-surface temperature over the tropical Atlantic. *Int. J. Climatol.,* **24,** 45–55.

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.

Roeckner, E., and Coauthors, 1992: Simulation of the present-day climate with the ECHAM model: Impact of model physics and resolution. Max Planck Institute for Meteorology Rep. 93, Hamburg, Germany, 171 pp.

——, and Coauthors, 1996: The atmospheric general circulation model ECHAM4: Model description and simulation of present-day climate. Max Planck Institute for Meteorology Rep. 218, 90 pp.

Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.,* **115,** 1606–1626.

——, and ——, 1989: Precipitation pattern associated with the high index phase of the Southern Oscillation. *J. Climate,* **2,** 268–284.

——, and ——, 1996: Quantifying Southern Oscillation–precipitation relationships. *J. Climate,* **9,** 1043–1059.

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

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Rowell, D. P., 1998: Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate,* **11,** 109–120.

Saha, S., and Coauthors, 2006: The NCEP climate forecast system. *J. Climate,* **19,** 3483–3517.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata, 1999: A dipole mode in the tropical Indian Ocean. *Nature,* **401,** 360–363.

Schmid, C., R. L. Molinari, R. Sabina, Y. H. Daneshzadeh, X. D. Xia, E. Forteza, and H. Q. Yang, 2007: The real-time data management system for Argo profiling float observations. *J. Atmos. Oceanic Technol.,* **24,** 1608–1628.

Schneider, E. K., 2002: Understanding differences between the equatorial Pacific as simulated by two coupled GCMs. *J. Climate,* **15,** 449–469.

Schubert, S. D., M. J. Suarez, P. J. Pegion, M. A. Kistler, and A. Kumar, 2002: Predictability of zonal means during boreal summer. *J. Climate,* **15,** 420–434.

Servain, J., I. Wainer, J. P. McCreary, and A. Dessier, 1999: Relationship between the equatorial and meridional modes of climatic variability in the tropical Atlantic. *Geophys. Res. Lett.,* **26,** 485–488.

Shukla, J., and J. L. Kinter III, 2006: Predictability of seasonal climate variations: A pedagogical view. *Predictability of Weather and Climate,* T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 306–341.

Stockdale, T. N., M. A. Balmaseda, and A. Vidard, 2006: Tropical Atlantic SST prediction with coupled ocean–atmosphere GCMs. *J. Climate,* **19,** 6047–6061.

Tebaldi, C., R. L. Smith, D. Hychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.,* **136,** 3933–3946.

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

——, A. G. Barnston, and A. W. Roberson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate,* **20,** 2210–2228.

Van den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus,* **46A,** 314–324.

——, 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 215 pp.

——, H. J. Huang, and Y. Fan, 2003: Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001. *J. Geophys. Res.,* **108,** 8617, doi:10.1029/2002JD003114.

Wang, G., O. Alves, D. Hudson, H. Hendon, G. Liu, and F. Tseitkin, 2008: SST skill assessment from the new POAMA-1.5 system. BMRC Research Letter 8, 43 pp.

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

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007a: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135,** 118–124.

——, ——, and ——, 2007b: Generalization of the discrete Brier and ranked probability skill scores for weighted multimodel ensemble forecasts. *Mon. Wea. Rev.,* **135,** 2778–2785.

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.

Winkler, R. L., and A. H. Murphy, 1968: ''Good'' probability assessors *J. Appl. Meteor.,* **7,** 751–758.

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.

Zebiak, S. E., 1993: Air–sea interaction in the equatorial Atlantic region. *J. Climate,* **6,** 1567–1586.