Diagnostic Verification of the IRI Net Assessment Forecasts, 1997–2000

D. S. WILKS AND C. M. GODFREY*

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

10 September 2001 and 3 January 2002

1. Introduction

The International Research Institute (IRI) for Climate Prediction produces operational outlooks for seasonal (3-month periods) average temperature and for total precipitation, at lead times of 0 and 3 months (Mason et al. 1999). These outlooks are probabilistic in nature and subjectively produced. During the 1997–2000 period considered here, two seasonal forecasts, at 0- and 3-month lead times, were produced 4 times per year. The forecast quantities are three-dimensional vectors specifying probabilities of temperature or precipitation outcomes falling in the lower (‘‘below normal’’), middle (‘‘near normal’’), and upper (‘‘above normal’’) thirds of the respective climatological distributions appropriate to particular seasons and locations; at global land (excluding Antarctica) and nearby ocean locations.

For precipitation, these IRI forecasts began in late 1997, and are available for October–November–December (OND) 1997 and January–February–March (JFM) 1998 onward, for the 0- and 3-month leads, respectively. The temperature forecasts began one season later, and are available from JFM 1998 and April–May–June (AMJ) 1998. The underlying information is from hand-drawn maps of ‘‘probability anomalies’’ (e.g., Mason et al. 1999, p. 1864), that were subsequently digitized to latitude–longitude grids appropriate to available verification data.

The precipitation forecasts are analyzed in the following after projection onto a global 2.5° × 2.5° grid, consistent with the format of the Xie and Arkin (1997) precipitation data, which begins in 1979. Similarly, the temperature forecasts were projected onto a global 2° × 2° grid to match the Ropelewski et al. (1985) temperature dataset. The climatological reference period, on the basis of which individual seasons were classified as being either below-, near-, or above-normal, was taken to be 1979–96 in order that the classification boundaries (the lower and upper terciles of each climatological distribution) do not involve years that are also part of the forecast data. Temperature and precipitation verification data were available through the OND 2000 season.

Because the verification grids have constant latitude–longitude increments, grid points were progressively thinned at the higher latitudes in order to approximate equal-area grids. Figure 1 shows the 2.5° × 2.5° verification grid for the precipitation forecasts. For some of the analyses the data have been stratified into the six broad geographic areas: North America, South America, Europe, Africa, Asia (excluding Southeast Asia and Indonesia), and Australia/western Pacific (including Southeast Asia and Indonesia). In regions and seasons where precipitation is sufficiently rare (<15% of the annual climatological precipitation in that season), no forecast is issued and instead that portion of the forecast map is designated ‘‘dry season.’’ These dry season precipitation forecasts are regarded as missing data in the present analysis. However, the numerous 1/3–1/3–1/3 (or ‘‘climatology’’) forecasts are regarded as valid, nonmissing data.

2. Scalar scores

The IRI net assessment forecasts are probabilistic in format, and should be evaluated accordingly. A first and gross look at their performance is provided by the ranked probability score (RPS, Epstein 1969; Murphy 1971; Wilks 1995), which is the average squared difference between the *cumulative* probability distributions of the forecasts f_k and observation variable o_k:

$$\mathrm{RPS} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{3}\left[\left(\sum_{k=1}^{m} f_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2. \qquad (1)$$

Here f_1, f_2, and f_3, are the forecast probabilities for the below-, near-, and above-normal temperature or precipitation outcomes; the observation variables o_k are indicators equal to 1 for the category in which the observation occurred, and 0 for categories in which the

[*] Current affiliation: School of Meteorology, University of Oklahoma, Norman, Oklahoma.

Corresponding author address: Dr. D. S. Wilks, Dept. of Earth and Atmospheric Sciences, Cornell University, 1113 Bradfield Hall, Ithaca, NY 14853.
E-mail: dsw5@cornell.edu

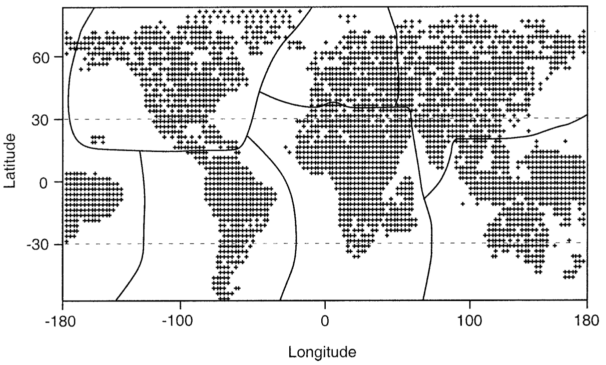

FIG. 1. The 2.5° × 2.5° precipitation grid of land and nearby ocean points, progressively thinned at higher latitudes to approximate an equal-area grid. Six continental divisions are indicated by solid lines.

TABLE 1. RPS skill scores, % [Eq. (2)], for the IRI temperature forecasts.

	0-month lead (JFM 1998–OND 2000)	3-month lead (AMJ 1998–OND 2000)
Global	8.9	4.2
Low latitudes ($\lvert\phi\rvert < 30°$)	17.8	10.5
High latitudes ($\lvert\phi\rvert > 30°$)	2.7	−0.3
Africa	25.6	9.1
Asia	1.1	1.3
Australia/west Pacific	19.8	13.0
Europe	5.0	2.7
North America	2.6	−1.3
South America	5.9	7.0

observation did not occur; and the average is over T space–time points. It is convenient to express results in terms of the RPS skill score:

$$\mathrm{SS}_{\mathrm{RPS}} = \frac{\mathrm{RPS} - \mathrm{RPS}_{\mathrm{clim}}}{0 - \mathrm{RPS}_{\mathrm{clim}}} \times 100\%, \tag{2}$$

where $\mathrm{RPS}_{\mathrm{clim}}$ indicates RPS for climatological forecasts ($f_1 = f_2 = f_3 = 1/3$) and 0 is the RPS for perfect forecasts.

Tables 1 and 2 show RPS skill scores [Eq. (2)] for the temperature and precipitation forecasts, respectively. On the basis of prior experience with verification of seasonal forecasts (e.g., Epstein 1988; Murphy and Huang 1991; Wilks 2000) it is not surprising that the temperature forecasts exhibit more skill than do the precipitation forecasts, and the forecasts made at 0-month lead are generally more skillful than those made 3 months in advance. There is a clear geographic dependence in the forecast skill, with nearly all of the global skill attributable to comparatively good performance at low latitudes (equatorward of ±30°), and essentially zero skill at higher latitudes in aggregate except for very modest positive performance for the 0-lead temperature forecasts. The disaggregation of skill scores according to the six geographic regions indicated by solid lines in Fig. 1 indicate essentially the same result: the predominantly low-latitude continents exhibit comparatively good skill while forecasts for the predominantly high-latitude areas are less successful.

3. Diagnostic verification

The results in Table 1 are useful as a first look, but as scalar summaries of an inherently multidimensional verification problem they are inevitably limited. In particular they do not indicate specific opportunities for potential improvements to forecast providers, and do not provide sufficient information to allow optimal use of the forecasts in decision making. A more complete approach is to adopt the perspective of ‘‘diagnostic verification,’’ which involves examining the joint frequency distribution of the forecasts and observations, $p(f_i, o_j)$, in order to diagnose particular strengths and weakness of a set of forecasts (Murphy and Winkler 1987, 1992).

It is convenient here to examine this joint distribution using the calibration-refinement factorization:

$$p(f_i, o_j) = q(o_j \mid f_i)\, r(f_i), \tag{3}$$

in which $q(o_j \mid f_i)$ denotes the set of conditional distributions (called the calibration distributions) for the observations given each of the possible forecasts, and the refinement distribution $r(f_i)$ expresses the frequency of use of each of the forecasts f_i. In the present case, each forecast consists of three related (because the three must sum to 1) probabilities, pertaining to the below-normal (cooler or drier than the 33d percentile of the climatological distribution), above-normal (warmer or wetter than the 67th percentile of the climatological distribution), and near-normal outcomes (between the 33d and 67th percentiles). Each of the three probabilities in a particular forecast can be regarded as pertaining to a pair of dichotomous events (so that there can be two values of o_j, $j = 0, 1$, corresponding, e.g., to below normal vs near- and above normal), which simplifies the verification analysis at the expense of producing some redundant information (Wilks 2000). The forecasts as issued then pertain to the ‘‘yes’’ outcome o_1. They are always integer multiples of 0.05, plus the climatological forecast 0.33, so each forecast probability will be one of the $i = 15$ values 0.05, 0.10, . . . , 0.30, 0.33,

TABLE 2. RPS skill scores, % [Eq. (2)], for the IRI precipitation forecasts.

	0-month lead (OND 1997–OND 2000)	3-month lead (JFM 1998–OND 2000)
Global	1.8	1.0
Low latitudes ($\lvert\phi\rvert < 30°$)	4.8	3.0
High latitudes ($\lvert\phi\rvert > 30°$)	−0.6	−0.6
Africa	3.2	0.8
Asia	−0.8	−0.1
Australia/west Pacific	6.7	5.8
Europe	−0.8	−1.2
North America	0.0	−0.7
South America	4.8	2.5

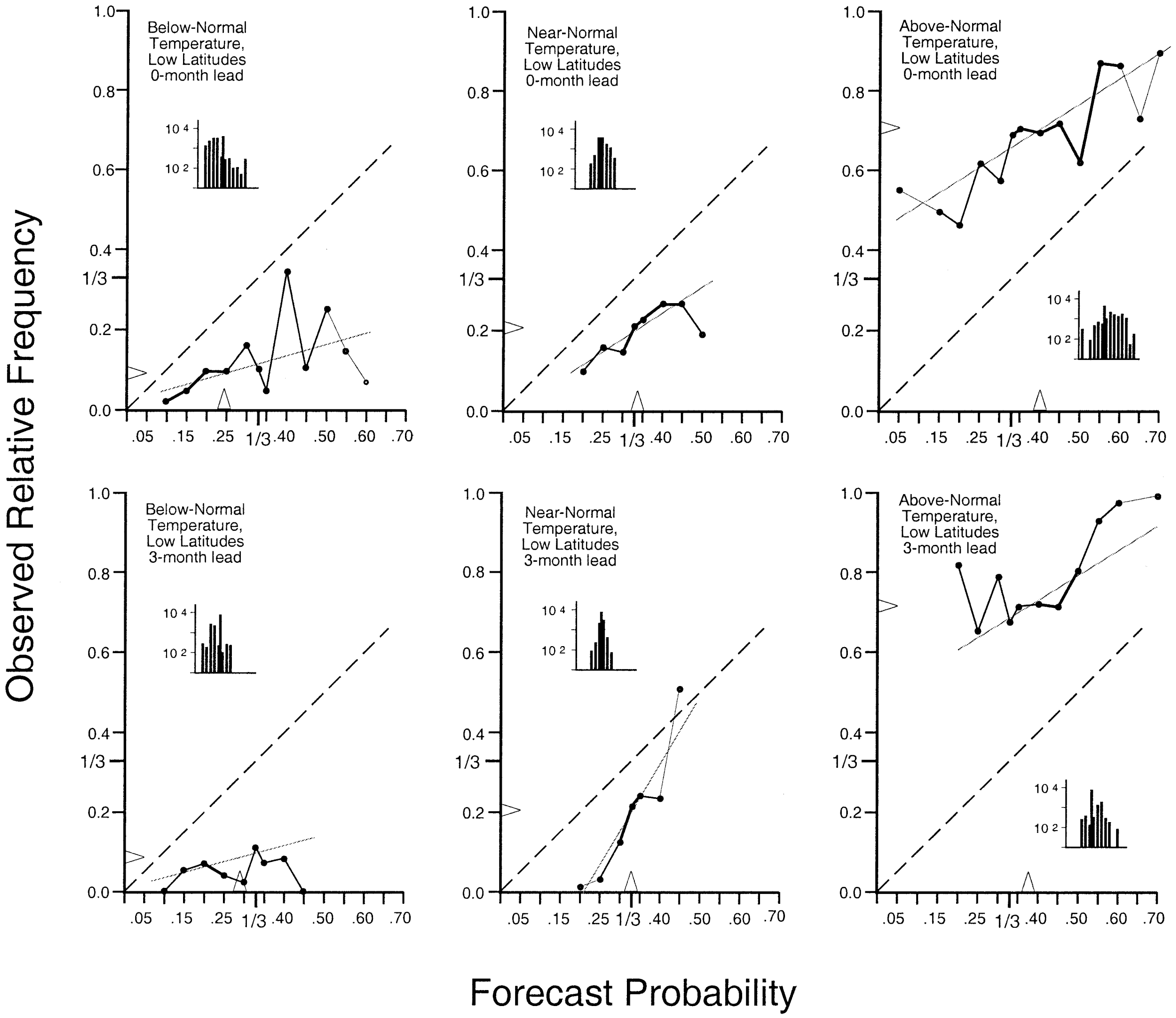

FIG. 2. Reliability diagrams [graphical depictions of Eq. (3)], for the temperature forecasts at low-latitude (equatorward of 30°) locations. Thickness of line segments connecting symbols defining the calibration functions $q(o_1 \mid f_i)$ increase with sample size, and the light lines show weighted least squares regressions summarizing each calibration function. Inset bar charts (note logarithmic vertical scales) portray the refinement distributions $r(f_i)$. Triangular symbols on the horizontal and vertical axes locate the average forecasts and average observations, respectively.

0.35, 0.40, . . ., 0.70. The factorization in Eq. (3) then consists of 15 conditional distributions $q(o_j \mid f_i)$ and a single refinement distribution $r(f_i)$, with distinct but related distributions q and r pertaining to the below-normal, near-normal, and above-normal outcomes.

In the present situation each of the refinement distributions $r(f_i)$ distributes probability among the $i = 15$ allowable forecasts, literally specifying the relative frequencies with which each of the 15 values have been used. Each refinement distribution can be interpreted as reflecting aggregate forecaster confidence: confident forecasts exhibit frequent and large departures from the climatological relative frequency (1/3), and hypothetical maximally confident forecasts would consist only of the ‘‘certain’’ probabilities 0.00 and 1.00. Forecasts exhibiting low confidence deviate rarely and quantitatively little from the climatological probability, and forecasts exhibiting no confidence consist of the climatological probability being forecast 100% of the time.

When looking at probability forecasts for dichotomous outcomes the calibration distributions $q(o_j \mid f_i)$ are Bernoulli distributions. Since each of these consists of 1 relative frequency, the set of 15 refinement distributions can be conveniently expressed graphically using reliability diagrams (e.g., Wilks 1995), in which the horizontal axis is forecast probability f_i and the vertical axis is $q(o_1 \mid f_i)$. A complete reliability diagram also includes a depiction of the frequency of use of the possible forecast values [i.e., $r(f_i)$], and thus is a full graphical representation of the joint distribution of the forecasts and corresponding observations [Eq. (3)].

Both unconditional biases (forecasts consistently too high or too low) and conditional biases (systematic forecaster over- or underconfidence) can be diagnosed from

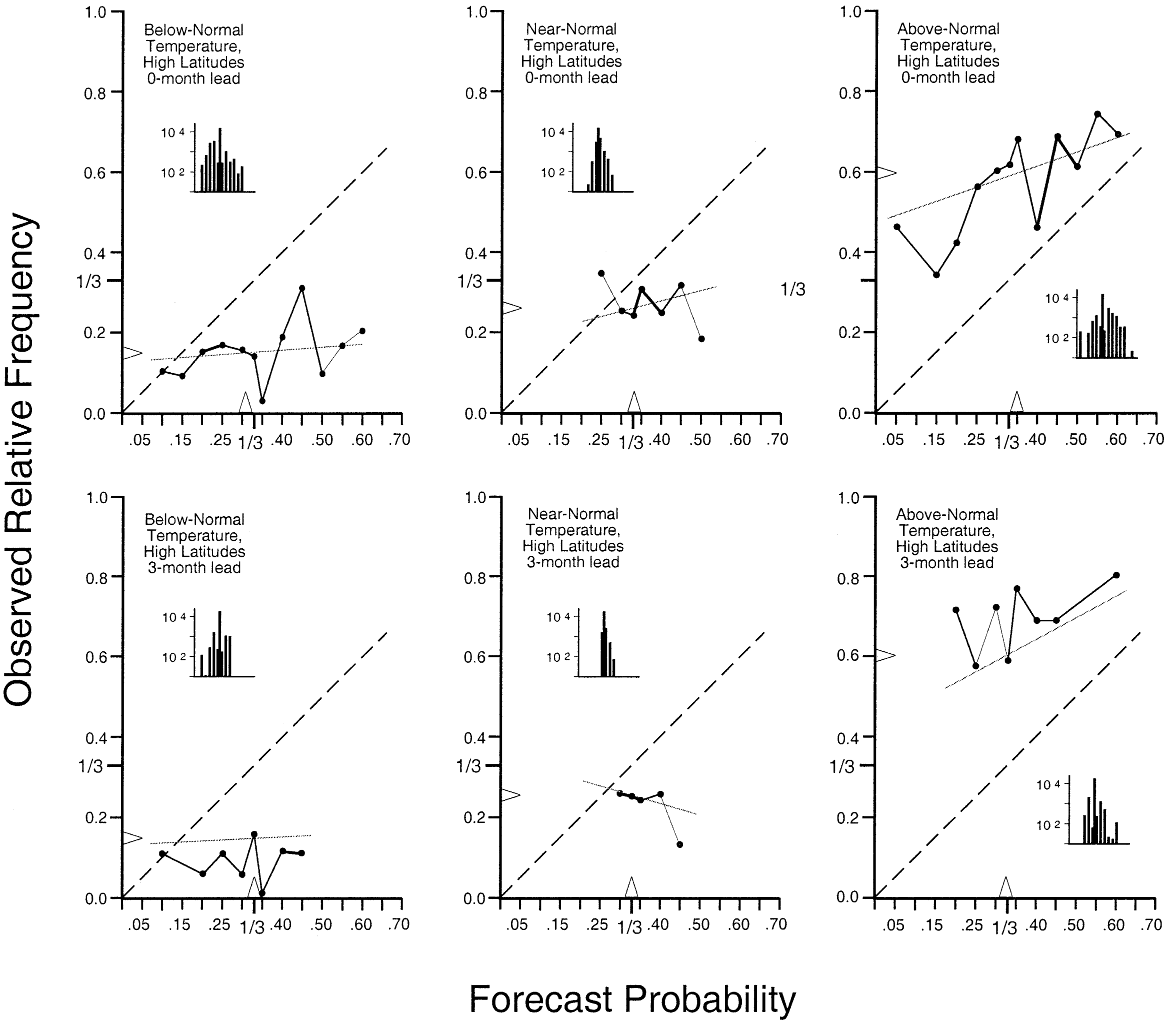

FIG. 3. As in Fig. 2, for high-latitude (poleward of 30°) temperature forecasts.

these plots (Wilks 2001; Wilks and Godfrey 2000). In particular, unbiased forecasts exhibiting an appropriate level of confidence produce reliability diagrams whose points fall close to the 1:1 line. Such forecasts "mean what they say," in the sense that $q(o_1 \mid f_i) \approx f_i$ for each i, given appropriate allowances for sampling variations. Biased forecasts exhibit points predominantly left (underforecasting) or right (overforecasting) of the 1:1 line. The relationship between $q(o_1 \mid f_i)$ and f_i for overconfident forecasts exhibits a slope shallower than the ideal 45°: extreme forecasts are not accompanied by extreme event relative frequencies, and correction of the problem would involve shifting mass in the refinement distribution $r(f_i)$ toward the climatological value (i.e., reducing forecaster confidence). Conversely, slopes steeper than 45° indicate underconfident forecasts, so that more frequent use of more extreme probabilities [increasing the dispersion of $r(f_i)$, or exhibiting greater confidence] would be justified.

4. Reliability diagram results

Reliability diagrams for the temperature forecasts are shown in Figs. 2 (low latitudes) and 3 (extratropics), separately for the below-normal, near-normal, and above-normal outcomes, and stratified according to the lead time. Figures 4 and 5 show the corresponding results for the precipitation forecasts. Thickness of the lines connecting points of the calibration functions $q(o_1 \mid f_i)$ increase according to the smaller sample size for each pair of points; with thickest lines connecting pairs of points both having $n > 1000$, medium lines connecting points with the smaller $n > 100$, and thinnest lines connecting points whose smaller sample size is less than 100. Points for which $n < 50$ are not shown in the main body of the diagram, but are indicated in the inset bar charts of the refinement distributions $r(f_i)$ (note log scale on vertical axes). The three contiguous bars in each refinement distribution identify the forecast probabilities $f = 0.30$, $f = 0.33$, and $f = 0.35$, and in

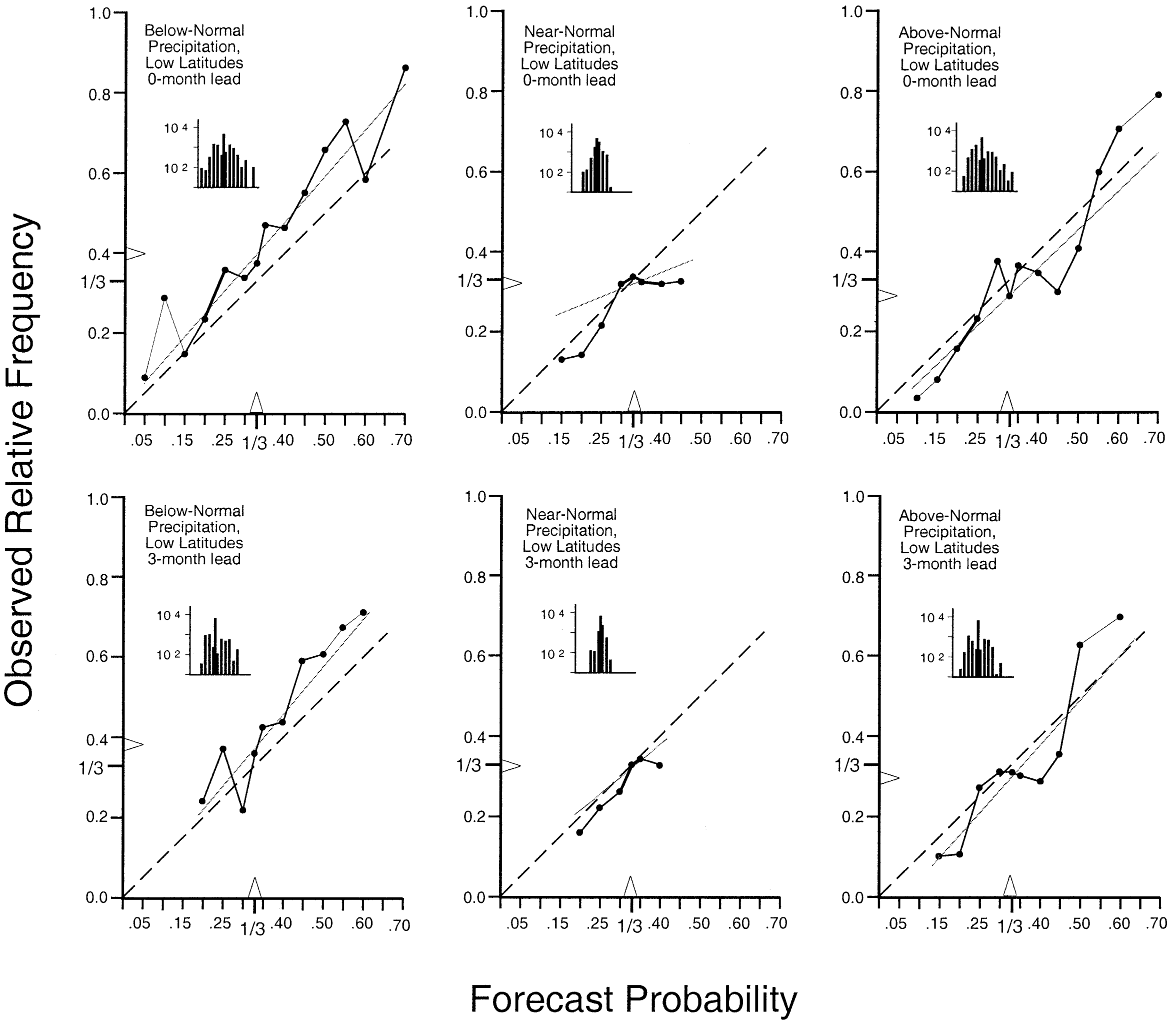

FIG. 4. As in Fig. 2, for low-latitude (equatorward of 30°) precipitation forecasts.

all cases the climatological $f = 0.33$ has been issued most frequently. The triangular symbols on the horizontal and vertical axes locate the average forecast and average observation, respectively. Light lines through the calibration functions are weighted least squares fits to the calibration points (Murphy and Wilks 1998) that in most cases help guide the eye in identifying the general character of the calibration functions.

The most prominent feature of the temperature verifications in Figs. 2 and 3 is the strong cold bias in the forecasts. The most recent few years have been warmer globally than the 1979–96 reference period used here (e.g., Lawrimore et al. 2001; Wigley 2000), and this is reflected in the relative frequencies of below-normal temperature outcomes being less than 0.2, and the relative frequencies of above-normal temperature outcomes being above approximately 0.6 (triangles on the vertical axes of Figs. 2 and 3). In contrast the average forecast in these cases are much closer to the climatological 1/3, indicating that these forecasts did not foresee the comparatively high temperatures during 1998–2000, in aggregate. This conclusion can also be reached on the basis of the evident underforecasting of the above-normal outcome (calibration points all to the left of the 1:1 lines) and overforecasting of the below- and near-normal outcomes (calibration points predominantly to the right of the 1:1 lines).

Aggregate forecast confidence for the low-latitude region (Fig. 2) is too high (calibration function slopes shallower than 45°), except for the 3-month lead, near-normal temperature outcome, where underconfidence is exhibited. Particularly for the above-normal outcome, the more extreme probabilities have been used comparatively frequently (the dispersion of the refinement distribution is comparatively high). The above-normal temperature forecasts for the extratropics (Fig. 3) exhibit similar calibration, although with a less confident refinement distribution, while the high-latitude forecasts for the below- and near-normal outcomes exhibit essentially no resolution (approximately flat calibration functions).

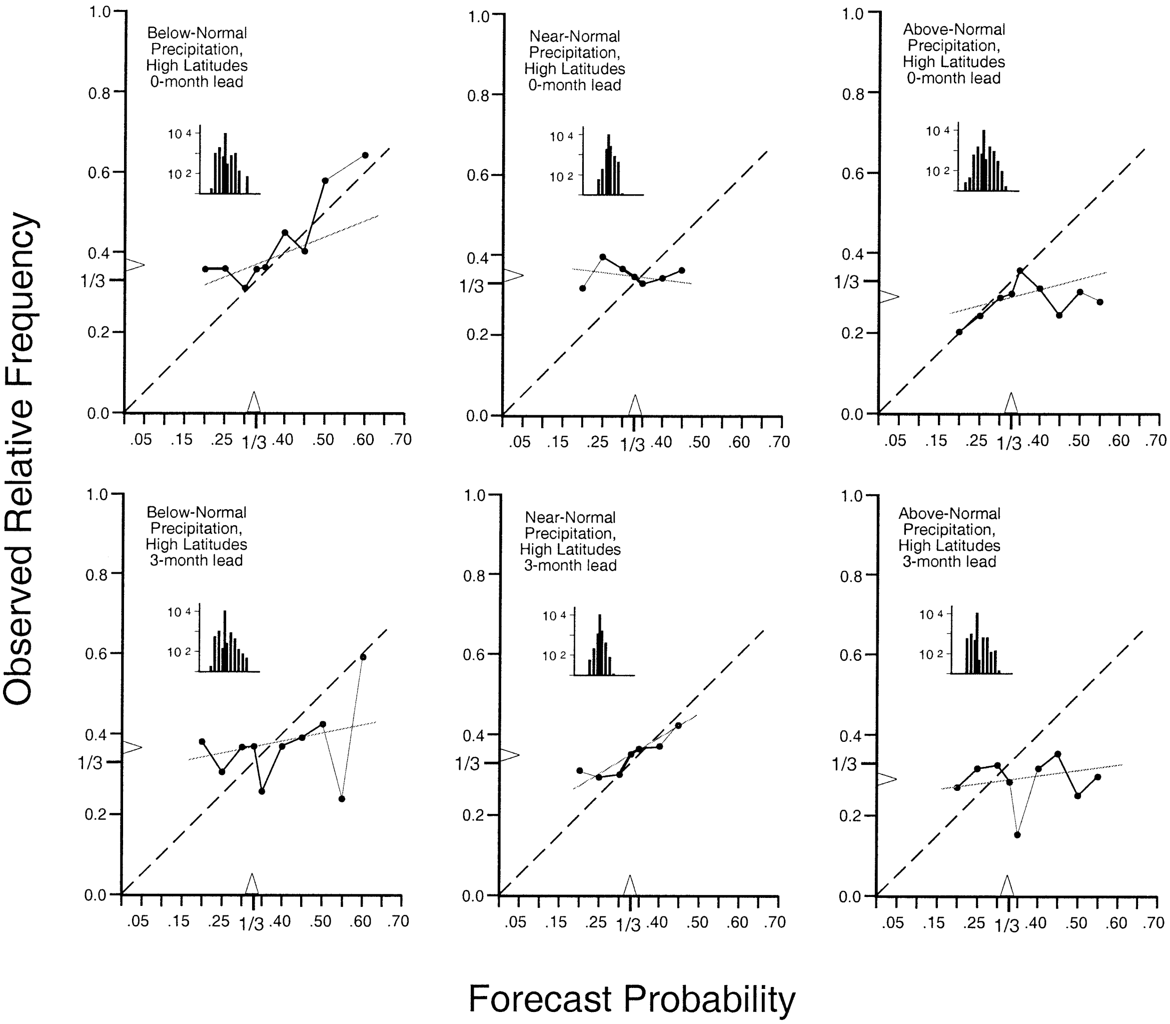

FIG. 5. As in Fig. 2, for high-latitude (poleward of 30°) precipitation forecasts.

Results for the precipitation forecasts (Figs. 4 and 5) also show bias, apparently related to 1998–2000 being relatively drier globally than 1979–96, although deviations of outcome relative frequencies from the climatological 1/3 are much less prominent than for temperature. Apart from the slight wet bias, precipitation forecasts for low latitudes (Fig. 4) exhibit good calibration overall, with the general slopes of the calibration functions being near 45°. The perhaps surprising (in light of the comparison between Figs. 2 and 4) result that the RPS skill score for low-latitude precipitation forecasts (Table 2) is smaller than for the corresponding temperature forecasts (Table 1) is a consequence of the precipitation forecasts being less confident (employing the more extreme probabilities less frequently). The exception to generally good calibration of the low-latitude precipitation forecasts is the near-normal outcome, where forecasts smaller than the climatological 1/3 are reasonably well calibrated, while forecasts above 1/3 do not resolve differences in the event outcomes. The former forecasts correspond to cases where either the wet or dry outcome is most probable, while the latter (unsuccessful) cases are those where the near-normal outcome was forecast as being the most probable. This feature was also evident in the globally aggregated results (not shown) and in preliminary results for 1997–99 (Wilks and Godfrey 2000).

The extratropical precipitation forecasts (Fig. 5) exhibit very little resolution, that is, the conditional relative frequencies $q(o_1 \mid f_i)$ are very near the average outcome (triangular symbols on the vertical axes), regardless of the forecast probabilities. These would seem to be of minimal utility for practical decision making.

5. Other data stratifications

The space required to display full reliability diagrams for all interesting stratifications of the data would be prohibitive, and those in Figs. 2–5 were chosen because the greatest differences in forecast performance relate

TABLE 3. Four-parameter summaries of the calibration-refinement factorization, following Murphy and Wilks (1998) for the IRI temperature forecasts. The parameters b_0 and b_1 are the intercept and slope, respectively, of the weighted least squares fits to the calibration functions $q(o_j \mid f_i)$, that are indicated as the light lines in Figs. 2–5. The parameters $\overline{f}$ and s_f are the mean and standard deviation, respectively, of the refinement distributions $r(f_i)$. Also included are the unconditional biases [Eq. (4)].

	0-month lead (JFM 1998–OND 2000)					3-month lead (AMJ 1998–OND 2000)				
(a) Cool outcome	b_0	b_1	$\overline{f}$	s_f	Bias	b_0	b_1	$\overline{f}$	s_f	Bias
Global	0.05	0.29	0.28	0.091	0.15	0.01	0.38	0.31	0.059	0.18
Low latitudes ($\lvert\phi\rvert < 30°$)	0.02	0.28	0.25	0.099	0.16	0.01	0.27	0.29	0.067	0.20
High latitudes ($\lvert\phi\rvert > 30°$)	0.12	0.07	0.31	0.074	0.16	0.13	0.05	0.33	0.044	0.18
Africa	−0.02	0.23	0.23	0.076	0.20	−0.01	0.11	0.30	0.059	0.27
Asia	0.18	−0.15	0.31	0.071	0.16	−0.02	0.49	0.33	0.041	0.19
Australia/west Pacific	−0.05	0.98	0.23	0.089	0.06	−0.17	1.25	0.28	0.072	0.10
Europe	0.04	0.24	0.30	0.073	0.19	−0.12	0.70	0.32	0.046	0.21
North America	0.13	0.11	0.31	0.084	0.15	0.27	−0.33	0.33	0.052	0.17
South America	0.10	0.04	0.29	0.120	0.18	0.10	0.06	0.29	0.072	0.18

	0-month lead (JFM 1998–OND 2000)					3-month lead (AMJ 1998–OND 2000)				
(b) Near-normal outcome	b_0	b_1	$\overline{f}$	s_f	Bias	b_0	b_1	$\overline{f}$	s_f	Bias
Global	0.11	0.42	0.34	0.041	0.09	−0.08	0.99	0.33	0.021	0.09
Low latitudes ($\lvert\phi\rvert < 30°$)	−0.17	0.64	0.34	0.053	0.14	−0.35	1.67	0.33	0.027	0.13
High latitudes ($\lvert\phi\rvert > 30°$)	0.17	0.26	0.33	0.029	0.08	0.35	−0.29	0.33	0.016	0.08
Africa	0.15	0.03	0.35	0.053	0.19	−0.46	1.82	0.33	0.022	0.19
Asia	0.42	−0.45	0.34	0.029	0.07	−0.10	1.06	0.33	0.015	0.08
Australia/west Pacific	−0.28	1.52	0.34	0.055	0.10	−0.13	1.06	0.34	0.030	0.11
Europe	0.27	−0.04	0.33	0.026	0.07	0.41	−0.46	0.33	0.018	0.08
North America	−0.03	0.83	0.33	0.027	0.09	−0.00	0.73	0.33	0.015	0.09
South America	−0.15	1.18	0.35	0.053	0.09	0.07	0.59	0.33	0.035	0.06

	0-month lead (JFM 1998–OND 2000)					3-month lead (AMJ 1998–OND 2000)				
(c) Warm outcome	b_0	b_1	$\overline{f}$	s_f	Bias	b_0	b_1	$\overline{f}$	s_f	Bias
Global	0.36	0.69	0.37	0.103	−0.25	0.32	0.89	0.35	0.066	−0.28
Low latitudes ($\lvert\phi\rvert < 30°$)	0.45	0.63	0.41	0.122	−0.30	0.48	0.62	0.38	0.081	−0.34
High latitudes ($\lvert\phi\rvert > 30°$)	0.47	0.35	0.35	0.080	−0.24	0.42	0.54	0.33	0.044	−0.27
Africa	0.60	0.52	0.42	0.097	−0.40	0.80	0.08	0.37	0.069	−0.46
Asia	0.65	−0.15	0.35	0.076	−0.25	0.05	1.67	0.33	0.043	−0.28
Australia/west Pacific	0.09	1.17	0.43	0.123	−0.17	−0.12	1.88	0.38	0.080	−0.21
Europe	0.57	0.14	0.36	0.075	−0.26	0.21	1.26	0.34	0.041	−0.30
North America	0.42	0.48	0.35	0.093	−0.24	0.57	0.06	0.33	0.055	−0.26
South America	0.39	0.66	0.36	0.140	−0.27	0.44	0.46	0.37	0.096	−0.24

to latitude (cf. Tables 1 and 2). However, the overall character of a reliability diagram can in most cases be captured using a few key statistics (Murphy and Wilks 1998). Tables 3 and 4 show five-parameter summaries of the reliability diagrams for temperature and precipitation forecasts, respectively; including global aggregation, latitude stratifications as in Figs. 2–5, and continental stratifications according to Fig. 1. Results for the latitude stratifications can be compared to Figs. 2–5 for perspective.

The two parameters b_0 and b_1 are the intercept and slope, respectively, of the weighted least squares lines through the calibration functions $q(o_1 \mid f_i)$. These serve to smooth sampling variations and summarize the dominant character of each calibration function, although as noted above, linear functions may not always be the best form for this purpose. Perfect forecasts would exhibit $b_0 = 0$ and $b_1 = 1$. The parameters $\overline{f}$ and s_f are the mean and standard deviation of the refinement distributions $r(f_i)$. The parameter s_f is a convenient index for aggregate forecast confidence, with more confident forecasts exhibiting higher standard deviations. The (unconditional) bias relates to the overall difference between the average forecast and the sample climatological relative frequency (i.e., the average observation indicated on the vertical axes of Figs. 2–5):

$$\text{bias} = \overline{f} - \overline{o} = (1 - b_1)\overline{f} - b_0. \qquad (4)$$

Positive bias thus indicates overforecasting, and negative bias indicates underforecasting.

Tables 3 and 4 show that the cold bias evident in Figs. 2 and 3, and the wet bias evident in Figs. 4 and 5, occur in all the geographic stratifications. The cool and near-normal temperature outcomes all exhibit overforecasting (positive bias), and the warm outcome is underforecast (negative bias). Similarly the dry precipitation outcome is underforecast, and the wet outcome is overforecast, for nearly all geographic groupings. As

TABLE 4. As in Table 3, for the IRI precipitation forecasts.

	0-month lead (OND 1997–OND 2000)					3-month lead (JFM 1998–OND 2000)				
(a) Dry outcome	b_0	b_1	$\bar{f}$	s_f	Bias	b_0	b_1	$\bar{f}$	s_f	Bias
Global	0.08	0.92	0.33	0.082	−0.05	0.10	0.81	0.33	0.064	−0.04
Low latitudes (|ϕ| < 30°)	0.02	1.15	0.33	0.103	−0.06	−0.01	1.17	0.33	0.078	−0.05
High latitudes (|ϕ| > 30°)	0.24	0.40	0.33	0.061	−0.04	0.30	0.20	0.33	0.052	−0.03
Africa	0.09	1.07	0.34	0.078	−0.11	0.20	0.75	0.33	0.059	−0.12
Asia	0.36	0.03	0.33	0.059	−0.04	0.26	0.31	0.34	0.050	−0.03
Australia/west Pacific	−0.03	1.17	0.33	0.127	−0.03	−0.20	1.52	0.34	0.094	0.02
Europe	0.21	0.37	0.32	0.059	−0.01	0.42	−0.26	0.32	0.043	−0.01
North America	0.21	0.50	0.32	0.064	−0.05	0.35	0.05	0.33	0.061	−0.04
South America	−0.05	1.34	0.32	0.101	−0.06	−0.07	1.33	0.33	0.075	−0.04
	0-month lead (OND 1997–OND 2000)					3-month lead (JFM 1998–OND 2000)				
(b) Near-normal outcome	b_0	b_1	$\bar{f}$	s_f	Bias	b_0	b_1	$\bar{f}$	s_f	Bias
Global	0.27	0.20	0.34	0.039	0.00	0.11	0.69	0.33	0.025	−0.01
Low latitudes (|ϕ| < 30°)	0.18	0.41	0.34	0.047	0.02	0.05	0.82	0.33	0.027	0.01
High latitudes (|ϕ| > 30°)	0.39	−0.12	0.34	0.031	−0.03	0.15	0.59	0.33	0.023	−0.02
Africa	0.42	−0.27	0.35	0.040	0.02	0.60	−0.77	0.34	0.020	−0.00
Asia	0.50	−0.42	0.34	0.029	−0.02	0.24	0.35	0.33	0.019	−0.02
Australia/west Pacific	0.12	0.52	0.33	0.056	0.04	−0.19	1.45	0.33	0.026	0.04
Europe	0.43	−0.19	0.34	0.032	−0.02	−0.06	1.31	0.33	0.024	−0.04
North America	0.33	0.01	0.33	0.033	−0.00	0.19	0.41	0.33	0.028	0.00
South America	−0.01	1.02	0.34	0.043	0.01	0.01	1.01	0.33	0.031	−0.02
	0-month lead (OND 1997–OND 2000)					3-month lead (JFM 1998–OND 2000)				
(c) Wet outcome	b_0	b_1	$\bar{f}$	s_f	Bias	b_0	b_1	$\bar{f}$	s_f	Bias
Global	0.04	0.74	0.33	0.081	0.04	0.06	0.70	0.33	0.062	0.04
Low latitudes (|ϕ| < 30°)	−0.03	0.96	0.33	0.102	0.04	−0.07	1.12	0.33	0.072	0.03
High latitudes (|ϕ| > 30°)	0.20	0.27	0.33	0.061	0.05	0.24	0.14	0.33	0.053	0.04
Africa	−0.05	0.88	0.31	0.082	0.09	−0.01	0.68	0.33	0.060	0.11
Asia	0.16	0.36	0.33	0.059	0.05	0.19	0.29	0.32	0.053	0.04
Australia/west Pacific	0.06	0.87	0.33	0.119	−0.02	−0.11	1.53	0.33	0.082	−0.07
Europe	0.27	0.08	0.33	0.056	0.03	0.32	−0.08	0.34	0.048	0.04
North America	0.14	0.45	0.34	0.070	0.04	0.24	0.19	0.33	0.060	0.03
South America	−0.04	0.97	0.33	0.097	0.05	−0.01	0.88	0.33	0.068	0.05

would be expected, confidence exhibited in the 0-lead forecasts is in all cases greater (larger standard deviation s_f) than for the 3-month lead forecasts. At the 0-month lead the confidence of the temperature forecasts is generally higher than confidence of the precipitation forecasts, although calibration function slopes b_1 that are shallower than 45° for the temperature forecasts (notably for continents other than Australia/Oceana) indicate that at least part of this greater confidence is misplaced. In all cases, aggregate forecast confidence for the low-latitude points is higher than for the high-latitude points.

6. Summary and conclusions

This paper has examined the performance of the IRI seasonal temperature and precipitation forecasts, 1997–2000, using both a summary scalar score (RPS skill score) and a comprehensive diagnostic verification. Not surprisingly, the RPS skill score indicates better performance for temperature forecasts than precipitation forecasts, and better performance for forecasts at 0-month lead than for forecasts at 3-month lead. It also points to dramatically better performance in the Tropics (equatorward of 30° latitude) than in the mid- and high latitudes.

The diagnostic verification indicates that these differences in aggregate RPS skill correspond in large part to differences in forecast confidence. For the low-latitude precipitation forecasts, the slopes of the calibration functions indicate that the level of confidence exhibited is appropriate. The temperature forecasts and the high-latitude precipitation forecasts exhibit overconfidence in general.

The most striking deficiency of these forecasts is the strong cold bias for temperature, which is evident for all locations and both lead times. Evidently this problem is related to the fact that the years considered were substantially warmer than the 1979–96 base period used here, so that the relative frequency of the ‘‘warm’’ outcome was approximately 0.6 (higher latitudes) to 0.7 (lower latitudes). In contrast, the average forecast for the above-normal outcomes was about 0.4 for the low-latitude points and very near the climatological 1/3 for the higher latitudes. This warm condition was a major feature of the global climate for 1998–2000 (the period covered by the temperature forecasts), but was evidently not anticipated by the IRI forecasts, in aggregate. Very similar problems have been seen in the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) forecasts for the United States, during 1995–98 (Wilks 2000). During this period the United States experienced much warmer and wetter conditions than during the 1961–90 reference period, that were also not foreseen in the CPC seasonal outlooks.

Note finally that these results are based on a rather small sample, including less than one full ENSO cycle. Indeed, during most of the period considered here, eastern Pacific sea surface temperatures were comparatively cold. Also, while the nominal sample sizes here are large, the ‘‘effectively independent’’ sample size is much smaller due both to the strong spatial correlation of seasonally averaged quantities, and the fact that contiguous regions having the same forecast can be quite large. Conclusions drawn here should therefore be regarded as tentative.

Acknowledgments. We thank Simon Mason and Tony Barnston of the IRI for supplying the digitized forecasts and corresponding observational data. This work was supported by the NOAA Economics and Human Dimensions program, under Grant NA86GP0555.

REFERENCES

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

——, 1988: Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting,* **3,** 69–75.

Lawrimore, J. H., and Coauthors, 2001: Climate assessment for 2000. *Bull. Amer. Meteor. Soc.,* **82,** S1–S55.

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

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, and J. Huang, 1991: On the quality of CAC’s probabilistic 30- and 90-day forecasts. Preprints, *16th Annual Climate Diagnostics Workshop,* Los Angeles, CA, Amer. Meteor. Soc., 390–399.

——, and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

——, and D. S. Wilks, 1998: A case study in the use of statistical models in forecast verification. *Wea. Forecasting,* **13,** 795–810.

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

Wigley, T. M. L., 2000: ENSO, volcanos and record-breaking temperatures. *Geophys. Res. Lett.,* **27,** 4101–4104.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.

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

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8,** 209–219.

——, and C. M. Godfrey, 2000: Diagnostic verification of the IRI Net Assessment precipitation forecasts, 1997–1999. *Proc. 25th Annual Climate Diagnostics and Prediction Workshop,* Palisades, NY, NOAA Climate Prediction Center, 153–156.

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