MULTIMODEL ENSEMBLING IN SEASONAL CLIMATE FORECASTING AT IRI

BY ANTHONY G. BARNSTON, SIMON J. MASON, LISA GODDARD,
DAVID G. DEWITT, AND STEPHEN E. ZEBIAK

The evolution of methods used by the IRI to combine predictions of several models into a single global climate forecast is described.

THE ORIGINAL IRI FORECAST SYSTEM. The International Research Institute (IRI) for Climate Prediction began issuing seasonal forecasts of global precipitation and temperature in late 1997, consisting of a forecast for the upcoming two consecutive 3-month periods (Mason et al. 1999a). Forecasts were issued quarterly, for the seasons of January–March, April–June, etc. The IRI's final issued forecasts have been called "Net Assessments." All of the IRI's forecasts, including those most current, can be found online at http://iri.columbia.edu/climate/forecast/net_asmt/.

As described in Mason et al. (1999a) and Goddard et al. (2003a), the IRI's approach to making forecasts has been a two-tiered process in which a prediction is first made for the sea surface temperature (SST) in the global oceans, and then the SST prediction is used as a driver of a forecast for the atmospheric climate. The climate forecasts are issued as probabilities of each of three equiprobable categories (above, near, and below normal with respect to a recent 30-yr period), based largely on a set of ensembles of dynamical atmospheric general circulation model (AGCM) predictions. A mix of dynamical and statistical models has been used to construct the SST predictions, varying by tropical ocean basin. These include (but are not limited to) the National Centers for Environmental Prediction (NCEP) coupled ocean–atmosphere dynamical model for predictions of the tropical Pacific SST (Ji et al. 1998), separate canonical correlation analysis (CCA) predictions for the tropical Atlantic and Indian Oceans, and damped persistence for the extratropical oceans with 3 months *e*-folding time. These predictions were smoothly blended at their geographical interfaces. For the first forecast season, in addition to such evolving SST predictions, the observed SST anomalies from the most recently completed calendar month were used as another SST prediction scenario—that is, SST anomaly persistence.

As a supplement to Goddard et al. (2003a) that details the skills of the IRI's climate forecasts during

AFFILIATIONS: BARNSTON, MASON, GODDARD, DEWITT, AND ZEBIAK—International Research Institute for Climate Prediction, Columbia University, Palisades, New York
CORRESPONDING AUTHOR: Anthony G. Barnston, International Research Institute for Climate Prediction, Columbia University, 61 Route 9W, Monell Rm. 227, Palisades, NY 10964
E-mail: tonyb@iri.columbia.edu
DOI: 10.1175/BAMS-84-12-1783

In final form 20 August 2003

their first four years, this paper discusses IRI's newly updated operational forecasting system. The new system automates the synthesis of the predictive outputs of several AGCMs, replacing more subjective human deliberation. Here we examine the implications of the newer system on forecast skill and reliability.

As discussed in Goddard et al. (2003a), while statistical tools played a role in formulating the IRI's climate forecasts, three dynamical AGCMs were used most heavily: the ECHAM3.6 model from Max-Planck-Institut für Meteorologie in Hamburg, Germany (Deutsches Klimarechenzentrum 1992; Roeckner et al. 1992); the National Center for Atmospheric Research (NCAR) Community Climate Model version 3.2 (CCM3.2) in Boulder, Colorado (Hack et al. 1998; Hurrell et al. 1998; Kiehl et al. 1998); and the NCEP Medium-Range Forecast model (MRF9) in Washington, D.C. (Kumar et al. 1996; Livezey et al. 1996). The predictions from these AGCMs were recalibrated to reduce systematic errors based on a long (>40 yr) history of the models' simulations using observed SST and compared with corresponding contemporaneous observations [an Atmospheric Model Intercomparison Project (AMIP) design; Gates 1992; Gates et al. 1999]. Ensembles of AGCM runs were used to calibrate probabilistic predictions, and the ensemble mean was used to recalibrate the central tendency of the predictions (Mason et al. 1999b).

The process of bringing together the three predictions was done subjectively by the forecasters, who examined maps of each AGCM's predictions along with accompanying hindcast skill maps, and weighted the AGCMs to form the combined forecast. Hence, the atmospheric predictions from the three AGCMs were generally weighted unequally, in proportion to their historical performance. For the first forecast season, the results using predicted SST versus persisted SST anomalies were also weighted subjectively by the forecasters in view of knowledge of the uncertainty of the predicted SST, based on their own experience and on formal studies relating to the given season and location.

A major disadvantage of such a system is the large time requirement in human resources. In 2001, forecasts began to be issued on a monthly instead of quarterly basis, and for four overlapping 3-month periods rather than two consecutive nonoverlapping periods. Because of the resulting sixfold increase in forecast production, automation of the process became imperative.

In this paper, implications for possible changes in forecast reliability and skill resulting from the automation of the IRI's Net Assessment forecast process are examined. In section 2 the more automated procedure is described, and in section 3 we provide a retrospective evaluation of the skill of the objective portion of the automated procedure for the 4 yr over which IRI forecasts have been issued operationally.

THE MORE AUTOMATED IRI FORECAST SYSTEM.

When monthly forecast issuance began in June 2001, new methods for automating forecast production also began, focusing mainly on combining the predictions of several AGCMs into a single forecast.

Objective multimodel ensembling. Two multimodel ensembling methods were implemented. One is Bayesian in nature (Rajagopalan et al. 2002; A. W. Robertson et al. 2003, unpublished manuscript, hereafter RLZG), and the other is a canonical variate technique (Mason and Mimmack 2002). Both methods estimate an optimum relative weighting of the individual AGCM predictions for a given season and location, based on the past performance of seasonal simulations for the period 1950–97 using observed SST To date, both have operated on an individual gridpoint basis. (For more information about these two techniques, see the sidebar.) Model weights are formulated in terms of AGCM performance with "perfectly" known SST anomalies, defining an upper limit of operational forecast skill in which knowledge of future SST is imperfect. The estimates of the forecast climate probability anomalies therefore still needed to be modified (usually weakened) subjectively by the forecasters, depending on the region and the perceived uncertainty of the associated predicted SST. The multimodel ensembling system carries an implicit assumption that the ratio of the weights among AGCMs would be, on average, approximately the same with imperfectly as with perfectly predicted SSTs. Here, "imperfectly" pertains to both evolving and persisted SST predictions, for which the same relative weighting across AGCMs is applied. Predictions resulting from the two SST prediction scenarios are currently weighted equally relative to one another, although in practice the evolving SST scenario receives greater weight because more AGCMs are run using it than are run using the persisted SST scenario.

The consolidated predictions produced by the two schemes were found to have moderately high spatial correlation, with the Bayesian method tending to produce probabilities deviating further from climatological probabilities than the canonical variates. Their

MORE ABOUT THE MULTIMODEL ENSEMBLING TECHNIQUES

In the Bayesian method, the weights are derived to optimize a likelihood score over the set of *n* hindcast years. The score is formulated by the product of the forecast probabilities that had been assigned to the observed tercile-based category. A "regularization" process accounts for the greater certainty associated with the performance of AGCMs having a larger effective sample size (e.g., longer historical record of simulations, or greater number of ensemble members). As described in Rajagopalan et al. (2002), an optimization algorithm known as Feasible Sequential Quadratic Programming (FSQP; Zhou and Tits 1993) is used to maximize the likelihood score of the combined forecasts. The optimization includes a baseline model that always issues the climatology forecast (33.3% probabilities for each tercile-based category), so that the final combined forecast includes an appropriate weight for climatology as well as for the AGCMs. In locations and seasons in which the AGCMs have relatively high skill, the climatology model exerts relatively little influence, while in lower-skill situations it would be assigned higher weight, thereby weakening AGCM predictions toward climatological probabilities. It is believed that the Bayesian method can consolidate the AGCM forecasts to produce higher skill than a simple average across AGCMs. While the Bayesian procedure does not explicitly account for redundancies among model forecasts, it does so implicitly through the optimization scheme, which would be unlikely to result in very high weights being assigned to both of two skillful but largely redundant models.

The canonical variate method (Huberty 1994; Mason and Mimmack 2002) involves constructing linear combinations of the ensemble predictions to maximize differences among the means of forecasts across the predictand categories (terciles of temperature or precipitation). The ensemble predictions are ranked by model in order from the highest to the lowest prediction value. Prior to the canonical variate analysis, an EOF analysis is performed, which produces modes representing shifts in the overall mean, the ensemble variances, or other broad features of the forecast distributions. Through the EOF analysis, as well as the main part of the procedure, the canonical variate method accounts for redundancies among model forecasts—a desirable characteristic missing in the subjective method used earlier by IRI. A summary of the formulation of canonical variates is provided in appendix C of Mason and Mimmack (2002).

hindcast skills differ from one another spatially and seasonally, but are similar overall. For simplicity the predictions resulting from the two schemes were averaged with equal weights. The resulting skill of retrospective hindcasts over the 4-yr period of the IRI's forecasts will be compared with that of the IRI's issued Net Assessment forecasts over the same period in section 3.

Expanded set of AGCMs. A benefit of using objective multimodel ensembling schemes is that the set of AGCMs included in the consolidation process can be changed with minimal human effort. One only needs to produce the multidecadal SST-based simulation history for any new AGCM. Thus, when the automated weighting system began in June 2001 it became easier to add AGCMs, and by October 2001 six AGCMs were used (Table 1). Figure 1 illustrates the IRI's more automated forecast system as of September 2003. The six AGCMs listed in Fig. 1 differ slightly from those shown in Table 1 in that ECHAM3.6 was no longer included, and an AGCM from the Experimental Climate Prediction Center (ECPC) at Scripps Institution of Oceanography was added.[1]

Higher spatial resolution in issued forecasts. One aspect of the automation was the replacement of time-consuming, human-drawn forecast maps with automatically produced GrADS-based forecast maps (see GrADS Web site: http://grads.iges.org/grads). Figure 2 shows a formerly issued hand-produced forecast map for African precipitation [for July–August–September (JAS) 2000], and Fig. 3 shows a comparable automated map for a more recent forecast. The earlier maps featured broad regions sharing the same forecast probability distribution, while the automated maps allow more detail to be retained—as much as the forecasters believe is warranted. The task of spatial smoothing, previously conducted subjectively, is now done more objectively by computer using a Cressman analysis (Cressman 1959) with an approximately 600-km-weighting radius. The resolution of the gridded forecast field is 2.5° for precipitation and 2.0° for temperature, matching the resolution of the verification data, which come from the Climate

[1] The ECPC model is a revised version of the AGCM earlier implemented at NOAA/NCEP (Kanamitsu et al. 2002), with some changes to the physics as described in Kanamitsu and Mo (2003).

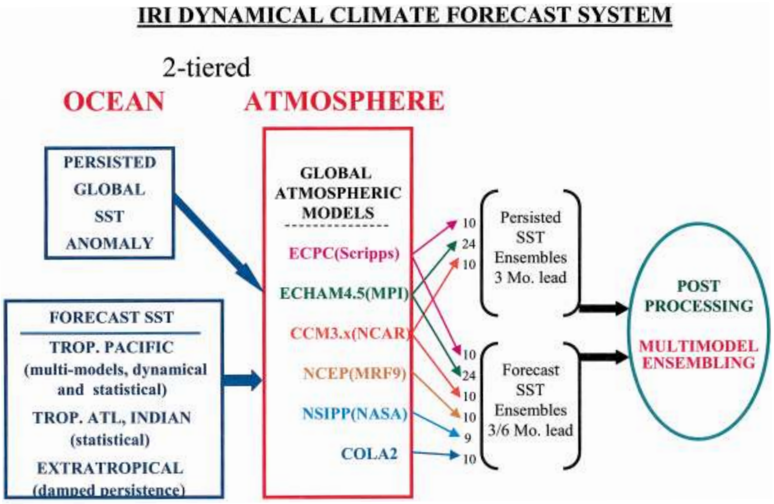

IRI DYNAMICAL CLIMATE FORECAST SYSTEM
2-tiered
OCEAN
ATMOSPHERE
PERSISTED GLOBAL SST ANOMALY
FORECAST SST
TROP. PACIFIC (multi-models, dynamical and statistical)
TROP. ATL, INDIAN (statistical)
EXTRATROPICAL (damped persistence)
GLOBAL ATMOSPHERIC MODELS
ECPC(Scripps)
ECHAM4.5(MPI)
CCM3.x(NCAR)
NCEP(MRF9)
NSIPP(NASA)
COLA2
10
24
10
Persisted SST Ensembles 3 Mo. lead
10
24
10
10
9
10
Forecast SST Ensembles 3/6 Mo. lead
POST PROCESSING
MULTIMODEL ENSEMBLING

Fig. 1. Schematic of the IRI's dynamical climate forecast system as of mid-2003. The flow begins with SST prediction, followed by ensemble runs of several AGCMs, followed by statistical postprocessing and consolidation into a single prediction using multimodel ensembling.

Anomaly Monitoring System (CAMS; Ropelewski et al. 1985) and from the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP; Xie and Arkin 1997), respectively. These resolutions do not differ greatly from that of most of the AGCMs, which is 2.8°.

SST scenario weighting. In addition to the approximately 10 ensemble runs produced from each of up to six AGCMs using an evolving SST prediction as the lower boundary condition, some of the AGCMs are also run with persisted SST anomalies for the shortest lead forecast season. Since there are unequal numbers of ensemble members between the two SST scenarios (not all models produce persisted SST anomaly predictions for logistical/institutional reasons), simple pooling with equal weights favors results using the evolving SST prediction. A more satisfactory automation of SST scenario weighting that not only accounts for unequal ensemble sizes, but that also explicitly accounts for differences in the skill of the AGCMs under specific SST scenarios, needs to be implemented. Currently, the forecasters must discuss the uncertainties in the two SST scenarios and subjectively weight the two sets of AGCM results accordingly. This involves estimating, from their own knowledge and experience, the likely differences in the climate effects attributable to features of the two SST scenarios.

RETROSPECTIVE SKILL OF MULTIMODEL ENSEMBLING OVER FOUR YEARS OF IRI FORECASTS. An estimate of the effect on forecast skill of using the multimodel ensembling procedure is provided by retrospectively examining the skills of the predictions issued over the 4-yr period of late 1997–2001, and comparing them with the skill of the actual, more subjectively derived IRI forecasts. A caveat, however, in addition to the obvious one of only having a 4-yr period available, is that the automated predictions differ not only in their lack of human judgment, but also in the lack of varying degrees of input from empirical tools such as ENSO-based probabilities (Mason and Goddard 2001) that were present in the subjective forecast procedure. Furthermore, the subjective forecast process evolved somewhat over the 4-yr period (Goddard et al. 2003a).

In this exercise, the ranked probability skill score (RPSS; Epstein 1969) is the main metric used, as it is an appropriate measure for probabilistically expressed forecasts. The RPSS gives increased credit not only for forecasting an enhanced probability for the observed category, but also for the strength of the probability assigned to that category. Similarly, it penalizes more severely for forecasts of a category not observed, if given with higher probability. The RPSS is designed to be 0 for the climatology forecast, and 1.0 (or 100%) for a forecast having a 100% probability for the category subsequently observed. It often ranges between 0 and 0.3 for climate forecasts whose probabilities deviate within ±20% of the climatological 33.3%, as is typically the case. The formula and computation of the RPSS are illustrated in appendix A of Goddard et al. (2003a).

Temperature. A comparison of skill for the original forecasts and the retrospective objective multimodel forecasts for temperature is shown in Fig. 4. It is apparent that in general, reliance upon the objective skill-weighted indications of the AGCMs would have delivered temperature forecasts having greater skill than that of the IRI's issued forecasts. Higher scores are seen in Indonesia, the Philippines, Japan, southern Africa, parts of the low-latitude Americas, and French Polynesia. Inspection of the geographical distributions of skill of the three individual AGCMs (see Figs. 6c–6e in Goddard et al. 2003a) indicates that these higher skill pockets were usually also achieved by one, two, or all three of the AGCMs, and did not come about specifically as a result of our multimodel ensembling. This is consistent with findings of others that, following the removal of biases in individual models, multimodel ensembling typically produces skills equal to or just slightly better than the skill of the most skillful AGCM in the set (Pavan and Doblas-Reyes 2000; Palmer et al. 2000; Peng et al. 2002).

A reason for the better performance of the multimodel ensemble predictions than the IRI's issued forecasts for temperature has to do with the disappointing performance of a statistical tool—the ENSO-based probabilistic composites—that contributed to the issued forecasts. (However, as discussed below, this same tool bolstered the skills of the precipitation forecasts.) Indeed, Goddard et al. (2003a) reported that the AGCMs were the highest scoring input tools to the IRI's temperature forecasts over the 4-yr period.

Figure 5 contains a set of reliability diagrams for three versions of the temperature forecasts over the 4-yr period: the IRI's Net Assessment forecasts, the objective multimodel ensemble, and one of the AGCMs. The concept of reliability is well developed

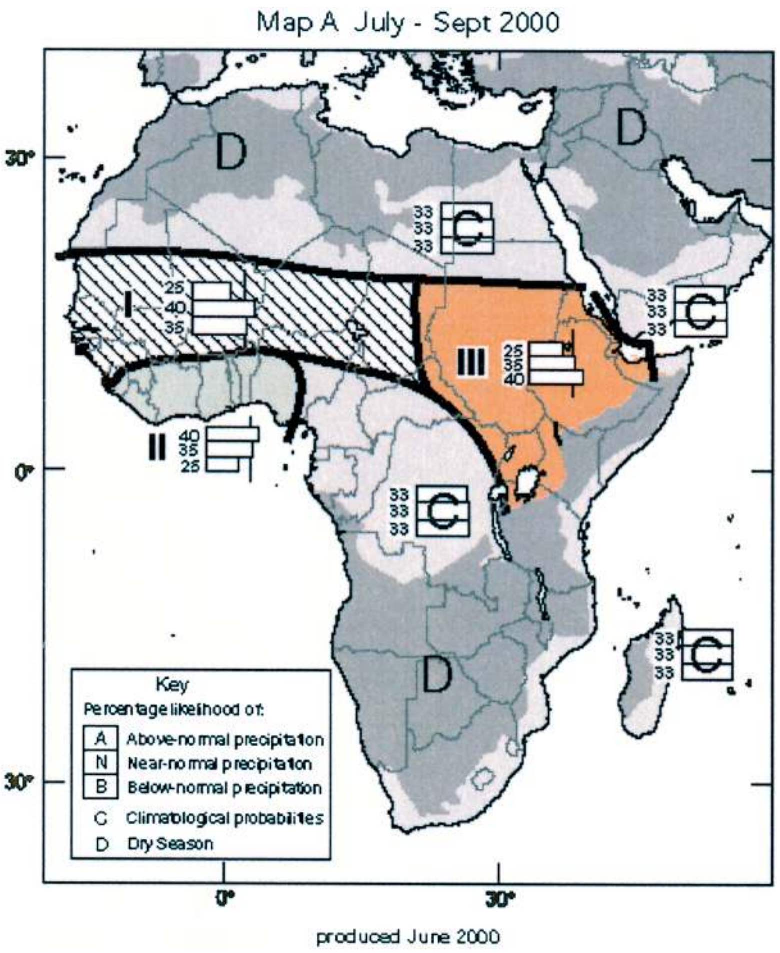

Map A July - Sept 2000
30°
0°
30°
0°
30°
D
D
D
33
33
33
C
I
25
40
35
II
40
35
25
III
25
35
40
Key
Percentage likelihood of:
A Above-normal precipitation
N Near-normal precipitation
B Below-normal precipitation
C Climatological probabilities
D Dry Season
produced June 2000

FIG. 2. Illustration of the IRI's forecast map graphic used before Jun 2001.

TABLE 1. The AGCMs used at IRI's forecast operation in Oct 2001, with their associated references. Revision of this list by Aug 2003 is indicated in Fig. 1.

Model	Where model was developed	Where model is run monthly
CCM 3.2	NCAR, Boulder, CO[a]	IRI, Palisades, NY
NCEP/MRF9	NCEP, Washington, DC[b]	QDNR, Queensland, Australia
ECHAM 3.6	Max Planck Institute, Hamburg, Germany[c]	IRI, Palisades, NY
ECHAM 4.5	Max Planck Institute, Hamburg, Germany[d]	IRI, Palisades, NY
NSIPP	NASA GSFC, Greenbelt, MD[e]	NASA GSFC, Greenbelt, MD
COLA	Center for Ocean–Land–Atmosphere Studies (COLA), Calverton, MD[f]	COLA, Calverton, MD

[a] Hack et al. (1998); Hurrell et al. (1998); Kiehl et al. (1998).

[b] Kumar et al. (1996); Livezey et al. (1996).

[c] Deutsches Klimarechenzentrum (1992); Roeckner et al. (1992); Goddard and Mason (2002).

[d] Roeckner et al. (1996).

[e] Bacmeister et al. (2000); Pegion et al. (2000); Schubert et al. (2002).

[f] Kinter et al. (1988); DeWitt (1996); Schneider (2002).

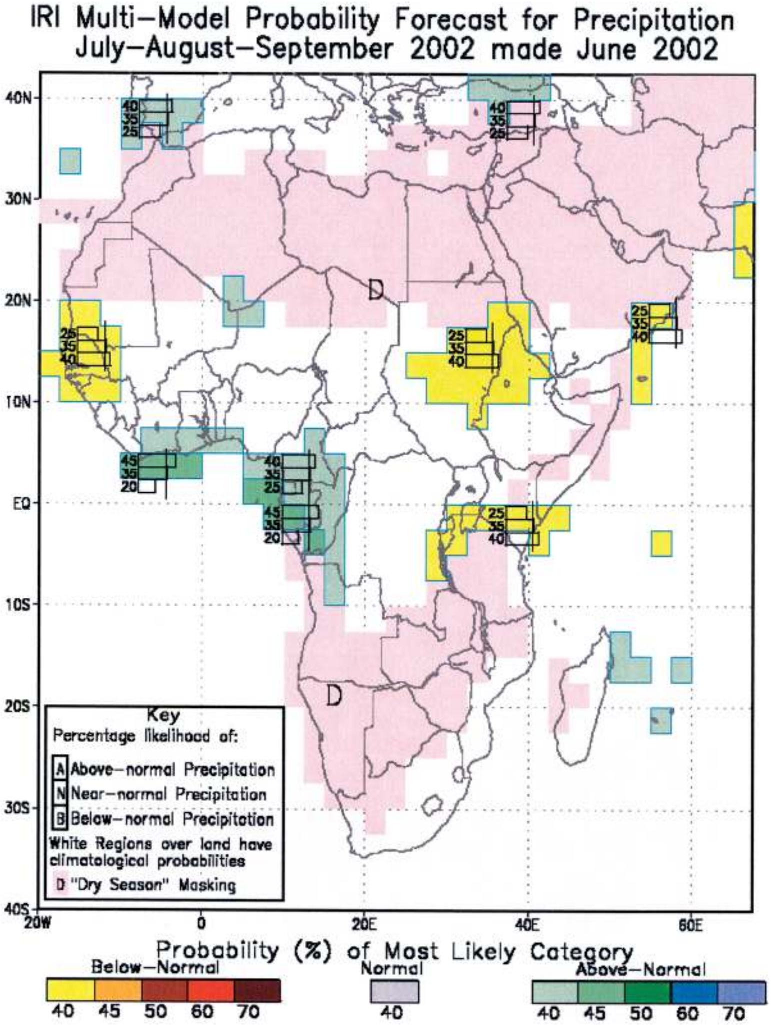

IRI Multi–Model Probability Forecast for Precipitation
July–August–September 2002 made June 2002
40N
30N
20N
10N
EQ
10S
20S
30S
40S
20W
0
20E
40E
60E
D
D
40
35
25
25
35
40
45
35
20
20
Key
Percentage likelihood of:
A Above–normal Precipitation
N Near–normal Precipitation
B Below–normal Precipitation
White Regions over land have climatological probabilities
D "Dry Season" Masking
Probability (%) of Most Likely Category
Below–Normal
40 45 50 60 70
Normal
40
Above–Normal
40 45 50 60 70

FIG. 3. Illustration of the IRI's forecast map graphic used since Jun 2001.

(e.g., Murphy 1973; Wilks 1995), and diagrams such as those shown in Fig. 5 were used in the evaluation of the IRI's Net Assessment forecasts (Wilks and Godfrey 2002) and the forecasts of the NOAA/CPC (Wilks 2000). Reliability refers to the correspondence between forecast probabilities for a given category (above, near, or below normal) and the relative frequencies of occurrence of the subsequent observations in the given category over a sufficient sample of cases. In such diagrams, the 45° dashed line represents perfect resolution and reliability, in which observed relative frequencies match forecast probabilities over the full range of forecast probabilities. Additional detail is provided in the caption of Fig. 5. The 4-yr period was strongly dominated by above-normal temperature as defined by the 30-yr normal in effect, as this category was observed more than 70% of the time (indicated by the green asterisk on the vertical axis) as opposed to the climatologically expected 33%.

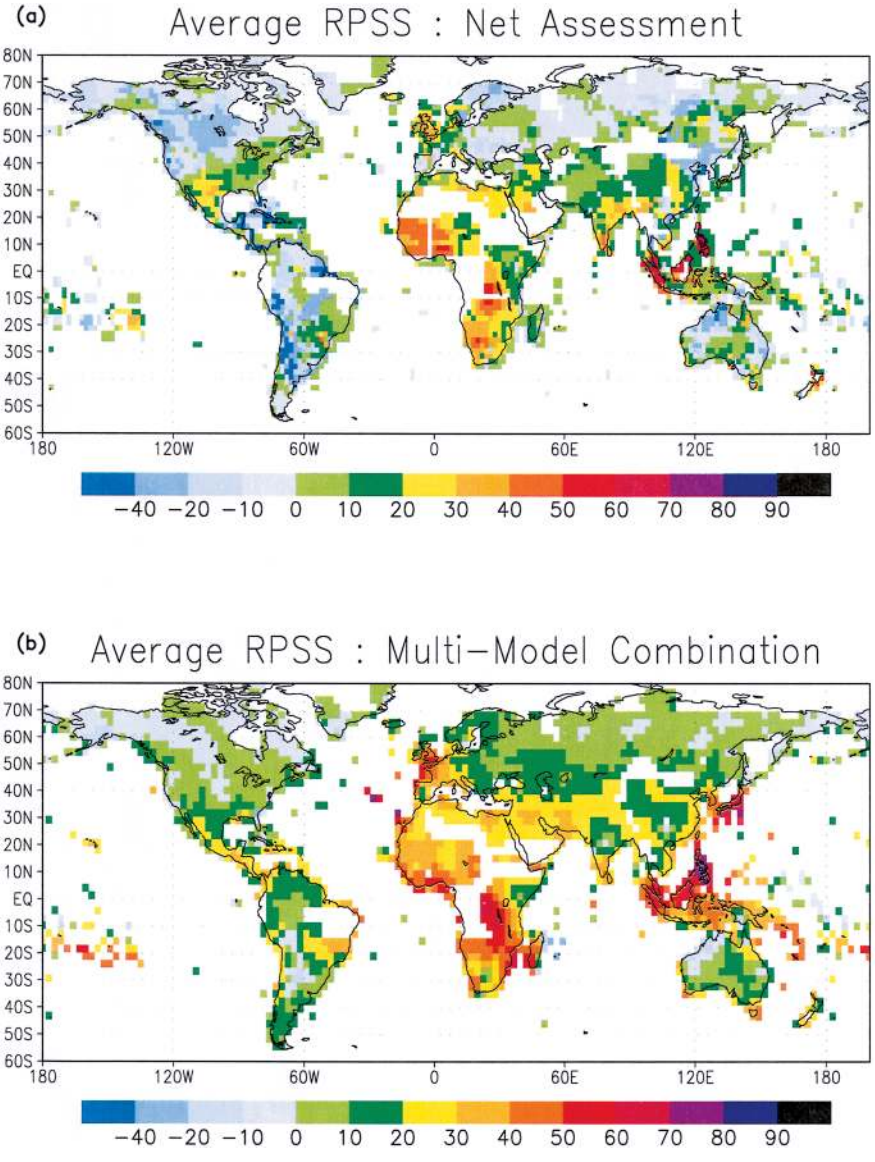
While both the issued IRI and the multimodel ensemble forecast sets gave considerably higher probabilities for above normal than expected long-term (roughly 40%–50%), they fell short of what occurred. This bias appears in the reliability diagrams as an overall vertical displacement of the curve from the 45° line, in which most of the points are located above (below) the 45° line in the diagram for above- (below) normal forecasts. The bias of the multimodel ensemble is somewhat smaller than that of the IRI's issued forecasts, particularly for below-normal temperature. In the multimodel combination plots, the blue curves and regression lines show results when forecasts of the three AGCMs are combined with equal weights. Equal weighting is seen to give results similar to those using the more sophisticated methods, except that the cool bias is reduced more using equal weights. More is said about this below.

When the slope of the reliability curve is shallower than that of the 45° line, overconfidence is indicated in the forecasts. Some overconfidence is present both in the probabilities in the IRI's issued forecasts and in the multimodel ensemble predictions, for forecasts for above- and below-normal categories.[2] In other words, increases in the forecast probability for a given category of temperature usually correspond to somewhat lesser increases in the probability of actually observing that category. An exception is seen in the multimodel predictions for below-normal temperature, but only for some of the high forecast probabilities.

As compared with the IRI's Net Assessment forecasts and the multimodel ensemble forecasts, the re-

[2] Reliability results for the near-normal category are not shown because little forecast skill was indicated for that category. This has been found repeatedly in previous studies (e.g., van den Dool and Toth 1991), and is related to the fact that overall shifting in the forecast probability distribution toward below or above normal is what is relatively most predictable, as opposed to a narrowing of the distribution, which is usually not sizeable. Large shifting can reduce the probability of the near-normal category, but changes probabilities in the two outer categories far more substantially.

liability diagrams for an individual AGCM (right column of Fig. 5) indicate a relatively milder underestimation (overestimation) of the observed frequency of the above- (below) normal temperature category—averaging near 60% (12%). That is, the AGCMs reproduced the overall strong tendency toward above-normal temperature relatively well. However, an extreme degree of overconfidence is revealed by a shallow slope for all three AGCMs. Confirming this overconfidence is a relatively flat frequency distribution of predicted probabilities (inset histogram), showing that large deviations from climatological forecast probabilities are common. This overconfidence suggests that 1) the shifts of the ensemble means away from climatology within an AGCM are too large, and/or 2) the spreads among ensemble members are too small. Either or both of these possibilities may be due partly to a lack of accounting for uncertainty associated with the models' own physical parameterizations (e.g., Toth and Vannitsem 2002; Palmer 2001). Model overconfidence is greatly reduced in the multimodel ensemble due to the averaging of the predictions across the models, as well as from spatial smoothing and the probabilistic damping in the two multimodel schemes based on the models' imperfect performance histories. A lack of accounting for the uncertainty in the predicted SST that forces the AGCM may also contribute to a tighter clustering in the AGCMs' climate predictions, and in the multimodel combination, than in the corresponding observations. An adjustment for this uncertainty was subjectively incorporated in the Net Assessment forecasts, and in principle could be objectively incorporated into the multimodel combination.


(a) Average RPSS : Net Assessment

(b) Average RPSS : Multi−Model Combination

Fig. 4. (a) Geographical distribution of RPSS (%) averaged over 16 IRI half-month lead seasonal Net Assessment forecasts of temperature for Jan–Feb–Mar 1998 through Oct–Nov–Dec 2001. Forecasts of the climatological distribution are included. Gray-shaded areas have insufficient verification data to be scored. [Note: This is the same as Fig. 2b in Goddard et al. (2003a)]. (b) As in (a), except for the objective two-scheme multimodel ensemble predictions using three AGCMs (ECHAM 3.6, CCM 3.2, and NCEP/MRF-9) as input, forced by predicted SST.

The RPSS of the individual AGCMs for temperature over the 4-yr period were unsurpassed by any other predictive tool or by the IRI's issued forecasts. This is largely due to the fact that their probabilities for above- (below) normal temperature averaged well above (below) their climatological values, as did the corresponding observed relative frequencies to an

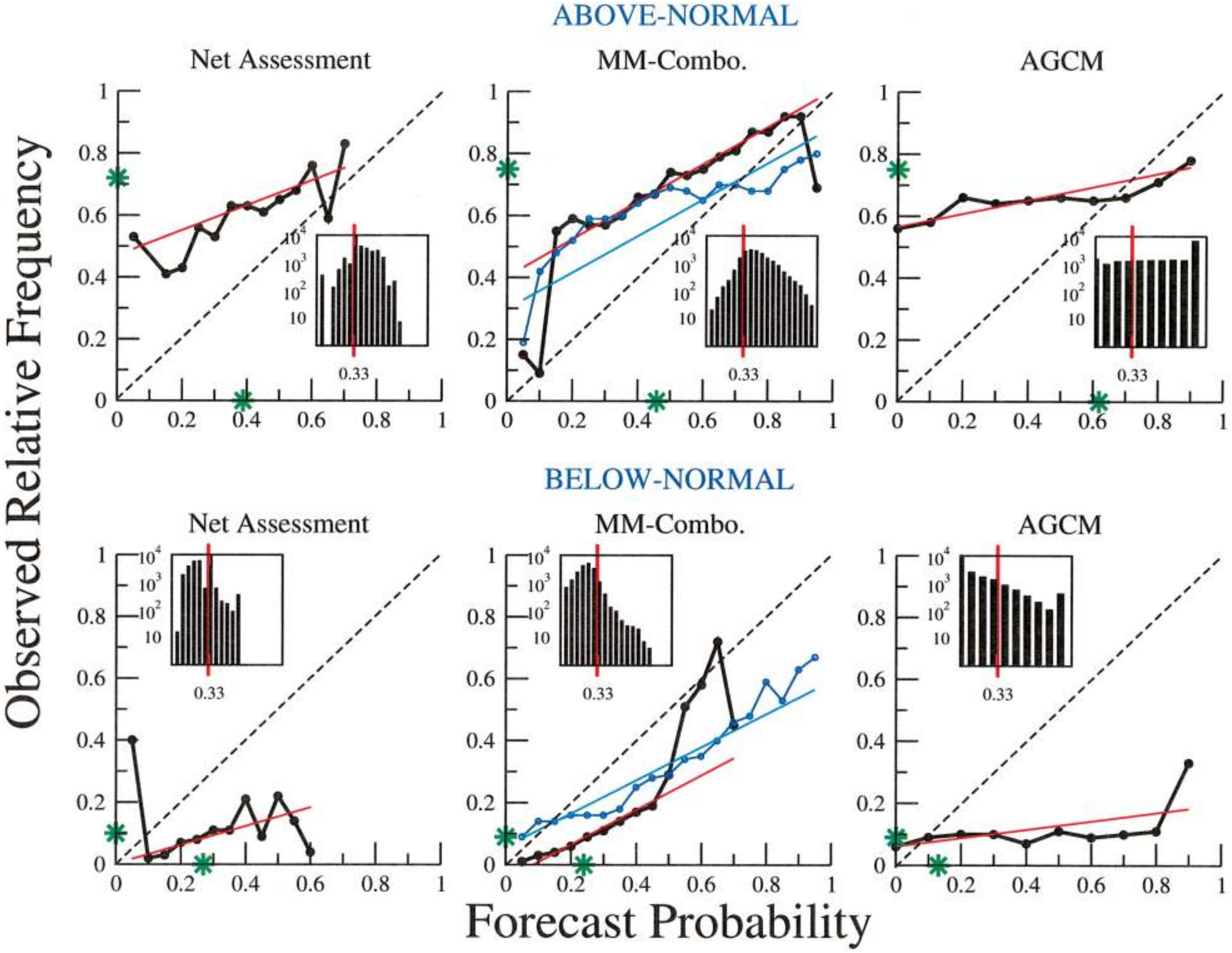


Fig. 5. Reliability diagrams for temperature forecasts at low latitude (30°N–30°S) over the 4-yr period for (left column) the IRI's issued forecasts (called Net Assessment), (middle column) the objective multimodel ensemble prediction, and (right column) for one of the three individual AGCMs. Just one of the AGCMs is included because all three of them have very similar reliability curves. (top row) Forecasts for above-normal temperature, and (bottom row) below-normal temperature. The *x* axis indicates forecast probability, and *y* axis relative observed frequency. The red line is a least-squares regression that takes into account the sample size represented by each point. The green asterisks on the axes show the overall mean of the forecast probabilities and observed relative frequencies. The inset histograms show the frequency with which each category of probability was forecast (with a logarithmic frequency scale), with the climatological probability (0.33) shown by the red vertical line. In the case of the multimodel combination (middle column), the blue dots and regression lines show results using a simple pooling (equal weighting) scheme as opposed to the two multimodel schemes. For IRI forecasts and multimodel predictions, the probability categories are centered on integer multiples of 0.05, while for the AGCM they are centered on integer multiples of 0.10.

even greater extent. This is indicative of good reliability in terms of the mean forecast probability shifts. However, the shallow slopes in Fig. 5 reflect poor forecast resolution (Murphy 1973): The relative frequency of above- or below-normal observed temperature is largely independent of the AGCMs' forecast probabilities. In Fig. 5 the multimodel ensemble forecasts (middle column) and issued IRI Net Assessment forecasts (left column) both have improved resolution as seen in the more favorable slopes and appropriately narrower ranges of forecast probabilities. However, because the latter two forecast sets have weaker mean probability shifts toward above-normal temperature, the positive contribution to their RPSS of the better resolution is outweighed by their greater cool bias.

Because the RPSS awards high credit for forecasts that probabilistically lean in the direction of the observations with high probabilities, the issuance of many predictions with higher probabilities for the above-normal category than what their long-term reliability-maximizing values would be (i.e., overconfidence) resulted in high RPSS scores over this very warm period. Such skill levels may not be sustainable over a longer period in which only moderately skillful but grossly overconfident predictions are made, and in which one of the two outer categories does not dominate the observations. This was in fact found to be the case in a multidecadal reliability analysis involving a set of AGCMs (Goddard et al. 2003b).

The above discussion points up the major role that multidecadal trends can play in climate diagnosis and prediction. Traditional 30-yr climatological base periods notwithstanding, use of a recent, much shorter base period (e.g., the last 10 yr) would serve the purpose of isolating interannual variability to simplify the interpretation of verifications such as those discussed above. Additionally, forecasts expressed with respect to the climatology of the last 10–12 yr would most likely be more meaningful to users.

Precipitation. A comparison of skill for the original forecasts and the retrospective objective multimodel forecasts for precipitation is shown in Fig. 6. In contrast to the results for temperature, skill for the multimodel ensemble is generally not more skillful than the IRI's issued forecasts, and falls short of them in some pockets of high skill such as over the Philippines and eastern Africa. The geographical distributions of skill of the three individual AGCMs (not shown) show patterns similar to that of the multimodel ensemble, but with even more sparsely spaced pockets of positive skill. Thus, all of the AGCMs had modest skills over much of the globe, and the multimodel ensemble made the most of a weak set of inputs. However, it should be noted that in the regions where precipitation is known to have some predictability, the predictability is usually limited to specific times of the year. Skills for these specific seasons alone, highlighted in Goddard et al. (2003a), are higher than those averaged over all seasons as in Fig. 6.

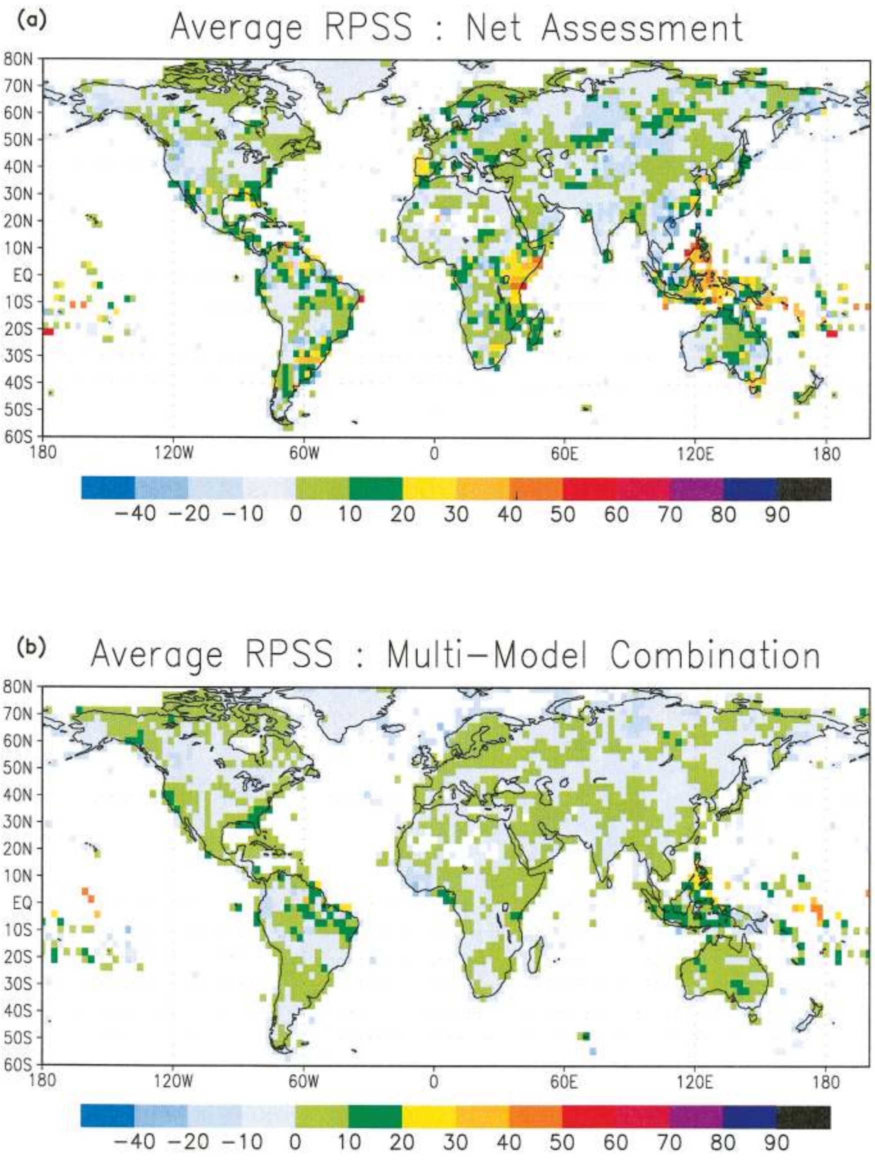


Fig. 6. (a) Geographical distribution of RPSS (%) averaged over 17 IRI half-month lead seasonal Net Assessment forecasts of precipitation for Oct–Nov–Dec 1997 through Oct–Nov–Dec 2001. Forecasts of the climatological distribution are included. Gray-shaded areas have insufficient verification data to be scored. [Note: This is the same as Fig. 3b in Goddard et al. (2003a)]. (b) As in (a), except for the objective two-scheme multimodel ensemble predictions using three AGCMs (ECHAM 3.6, CCM 3.2, and NCEP/MRF-9) as input, forced by predicted SST.

One reason that the objective multimodel ensemble predictions did not have greater skill than the more subjectively derived IRI forecasts is that, unlike the case for temperature, the supplementary inputs to the IRI's forecasts proved beneficial. In particular, the empirical ENSO-associated precipitation probabilities enhanced the skills of the IRI's precipitation forecasts in a number of regions having known sensitivities to ENSO. The AGCMs showed skill in predicting precipitation impacts to ENSO in Indonesia, the Philippines, most tropical Pacific islands, the southeastern United States, and small parts of eastern Africa and northeastern Brazil. The empirical ENSO-associated probabilities were fairly successful in many (but not all) of these regions, as well as some others such as southeastern Brazil and Uruguay, southern Africa, and northeastern Africa/Saudi Arabia. A combination of the two types of tools produced the most skillful result for precipitation, as was done in the subjective process leading to the IRI's issued Net Assessment forecasts. If the best possible forecasts are to be issued in a more objective, automated way, the process may benefit from including empirically based tools in addition to the AGCMs.

A possible reason that the AGCMs sometimes fell short of the empirical ENSO-associated probabilities is the presence of systematic errors in the AGCMs. While biases in mean and variance are effectively reduced on an individual gridpoint basis in the multimodel ensembling scheme, spatial displacement errors were not treated on a pattern level using multivariate statistical techniques such as CCA or singular value decomposition (SVD; e.g., Feddersen et al. 1999; Ward and Navarra 1997), or other techniques. Even small spatial shifts may negatively impact the precipitation skills of the AGCMs, since, in contrast to temperature, precipitation patterns often involve small-scale features (e.g., Gong et al. 2003). While spatial displacement errors are not a general or severe problem in AGCMs, they are identifiable. Spatial corrections have been found to increase the value of the AGCM predictions in particular regions such as southwest Asia (Tippett et al. 2003) and southern Africa (Landman and Goddard 2002), and work is under way to apply this process to IRI's AGCM predictions more regularly, where beneficial.

Figure 7 contains reliability diagrams presented in the same format as those shown in Fig. 5 for temperature forecasts. The 4-yr period was characterized by a slight preponderance of drier-than-normal conditions (40%–45%). This came mainly at the expense of the near-normal category rather than the above-normal category, as the latter was observed at least 30% of the time.[3] The IRI's issued forecasts, and particularly the multimodel ensemble predictions, did not reflect this dry tendency, and thus had wet biases with respect to the below-normal category as noted by most points being above the reliability-maximizing 45° line. The slope of the curve, being roughly similar to that of the 45° curve, reveals acceptable resolution (lack of overconfidence or underconfidence) in both the IRI's issued forecasts and in the multimodel ensemble predictions. This implies that increases in the forecast probability for a given category of precipitation correspond approximately correctly to comparable increases in the probability of observing that category. Results for the equal weighting version of the multimodel combination (blue curves) show greater overconfidence (i.e., worse resolution), although the bias of underestimating dryness is reduced.

The reliability curves for one of the AGCMs, representing the curves of the other two AGCMs also, reveals strongly overconfident predictions. This overconfidence, as revealed by the shallow slope of the reliability curve and the flat frequency distribution of forecast probabilities (inset histogram), was also noted above for temperature predictions. The overconfidence is largely eliminated in the multimodel ensemble through averaging across models, spatial smoothing, and the calibrating effects (including amplitude damping) of the filtering and weighting processes of the two multimodel schemes. As noted by the blue curves and regression lines in the middle column of Fig. 7, a simple equal-weighting scheme to combine the forecasts of the three AGCMs does not sufficiently reduce the overconfidence.

SUMMARY AND DISCUSSION: EFFECTIVENESS OF THE MULTIMODEL ENSEMBLE PROCESS.

The results above suggest that the automated, objective multimodel ensemble procedure effectively consolidates predictions from several AGCMs by weighting them using two formulations, each based on their historical performance under conditions of perfectly known SST. The procedure appears to successfully substitute for human re-

[3] This may be due to a difference in climatologies between the University of East Anglia data (New et al. 1999, 2000) used to define the terciles, and the CMAP data (Xie and Arkin 1997) used for verification. A check for relative biases during the overlap period did not reveal a problem, but differences in the tercile boundaries may exist. Another possibility is that the distribution of precipitation may in fact be changing toward more general dryness but with a higher frequency of extreme events, and therefore occasional very high seasonal rainfall totals (Dai et al. 1998; Karl et al. 1995).

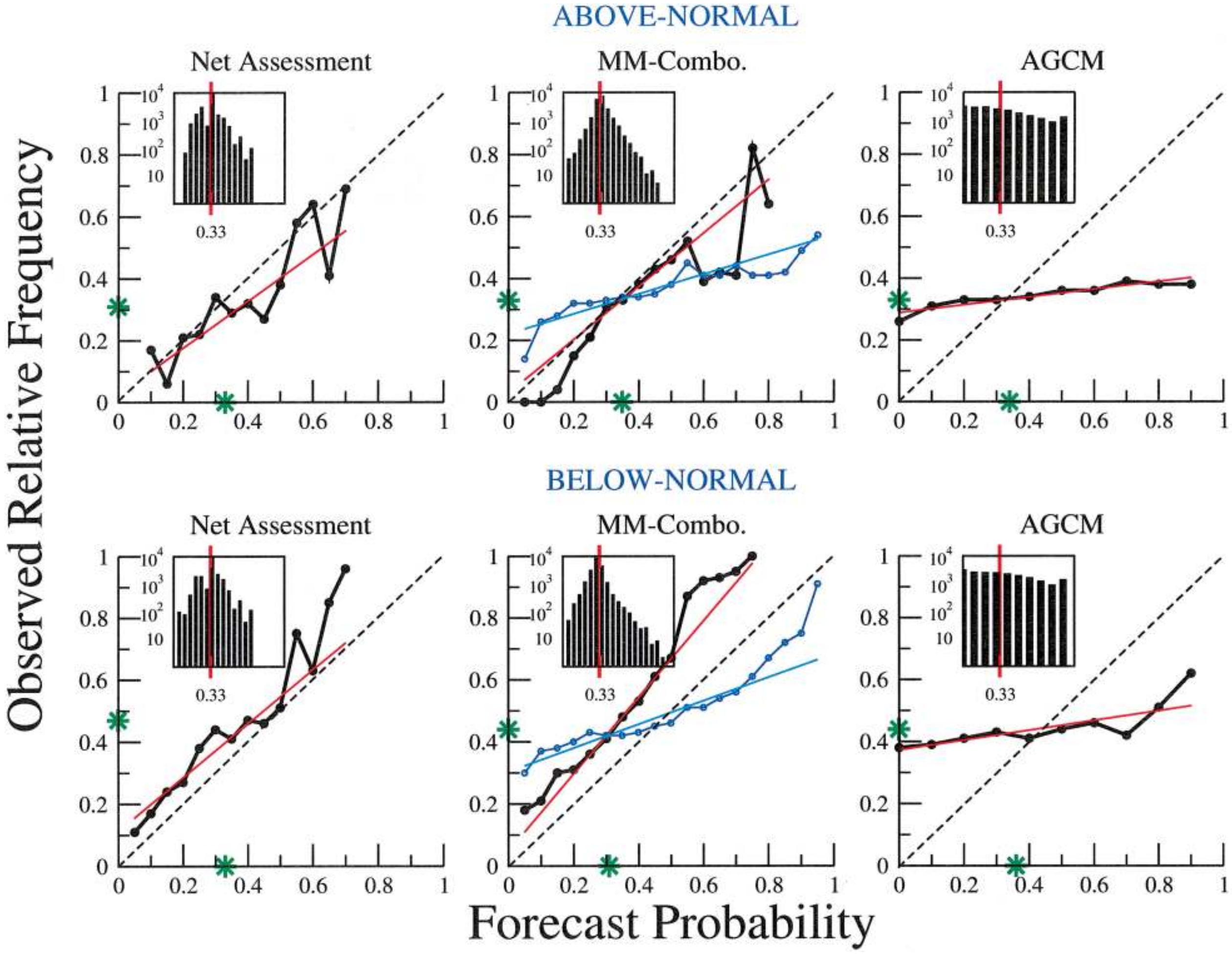


Fig. 7. As in Fig. 5 (reliability diagrams), except for precipitation forecasts. See the caption of Fig. 5 for more detail.

sources, in that the weighting and spatial smoothing is of approximately equal quality as when done subjectively.[4] The skill results for precipitation suggest that additional empirical tools could still beneficially complement the best combination of AGCMs. The marginally lower overall precipitation skill of the AGCMs alone compared with the skill of the AGCMs complemented by the statistical ENSO tool may be attributed to 1) model-specific systematic errors related to their physical parameterizations and/or 2) systematic errors in pattern shape and/or location, which were not statistically corrected in this study. In terms of reliability (correspondence between forecast probabilities and observed relative frequencies), the individual AGCMs are seen to be greatly overconfident for both temperature and precipitation. Overall biases in the AGCMs are not severe, despite some underestimation of the pervasive warmth and the tendency toward dryness in the 4-yr study period. The multimodel ensemble and the IRI's issued Net Assessment forecasts greatly reduced the overconfidence of the AGCMs, due to the conditional acceptance and recalibration of the model predictions, which can be done either subjectively by the forecasters or objectively and automatically using the multimodel ensemble schemes.

The multimodel ensemble and the IRI's issued forecasts did not improve upon the overall bias characteristics of the AGCMs over the 4-yr period—the underprediction of warmth and dryness. The three AGCMs had already been corrected for biases over a longer historical period, as the tercile boundaries were derived in terms of their own historical distribution. Underprediction of the warm conditions, which may

[4] It is impossible to assess this precisely, due to the inclusion of additional inputs to the IRI's issued forecasts. These empirical inputs appear to have increased the skill of the precipitation forecasts, and degraded the skill of the temperature forecasts over the 4-yr study period.

be regarded as a bias, can also be seen as an individual forecast underestimation in the sense that the warmth in many locations occurred more or less continuously throughout the period. This points to the fact that a 4-yr period is insufficient to perform a conclusive analysis, particularly when a variable does not have sufficient opportunity to span its normally expected range. A differing interpretation is that it was not the short period, but rather the decadally upward-trended temperature with respect to an "outdated" normal, that led to the preponderance of above-normal temperature that the AGCMs reproduced to a moderate extent, but not fully.

Although the two multimodel ensembling schemes exhibited somewhat differing characteristics, their skills and reliability diagrams (not shown) were comparable, and neither scheme was found to have undesirable characteristics. Because of the imperfectly correlated outputs of the two schemes, together they delivered a well-balanced product in terms of pattern and strength. It is worth noting that a simple method to combine the forecasts of the three AGCMs using equal weights (blue curves and lines in the middle column of Figs. 5 and 7) yielded results comparable to those of the two "smarter" schemes for temperature (Fig. 5)—in fact, with a better cool bias reduction. For precipitation, however, equal weighting yielded less effective correction of overconfidence than the two schemes, although it did slightly better with bias (Fig. 7). The Bayesian scheme was found to outperform equal weighting over a multidecadal period (Rajagopalan et al. 2002; RLZG). While its benefit is less clear in this 4-yr period, we cannot conclude that equal weighting would be as effective as the more sophisticated schemes.

It is possible to include one or more empirical tools in an objective, automated fashion as additional input to objective predictions of either precipitation or temperature. Empirical predictions could include, for example, the ENSO-related probabilistic composites already discussed (Mason and Goddard 2001), a CCA or SVD relating observed patterns of climate over land to patterns of SST anomalies (e.g., Barnett and Preisendorfer 1987; Graham et al. 1987; Wallace et al. 1992; Barnston and Smith 1996), and a decadal-scale persistence tool (e.g., "optimum climate normals") as used for forecasts for the United States by the NOAA/NCEP/CPC (Huang et al. 1996). Statistical spatial correction of the AGCM predictions [i.e., model output statistics (MOS)] would give the AGCM tools greater value. In sum, it is possible that the skill of IRI's final forecasts can be improved beyond what has been shown here.

A final element that may have helped the subjectively produced precipitation forecasts but not the automated forecasts is the ability of the forecasters to judge the reality of the evolving versus the persisted SST prediction in certain critical precipitation-determining seasons and regions, and to beneficially adjust the climate forecasts accordingly. Such deliberation was less frequently applied to the temperature forecasts. This factor is difficult to automate because it is related to uncertainties in the SST prediction—a prediction accepted as a fixed quantity in the forecast system and in this study.

The prediction of SST is a challenge in its own right, in which much progress is needed. Improvements in a global ocean observing capability should make possible new research. The introduction of new three-dimensional global upper-ocean data with the Argo data network [Wilson 2000; Commonwealth Scientific and Industrial Research Organisation (CSIRO) 2002], and the Pilot Research Moored Array in the Tropical Atlantic (PIRATA) array in the tropical Atlantic beginning in 1997 (Servain et al. 1998), may lead to improvements in SST prediction, particularly outside of the tropical Pacific where prediction skill is weakest. A more methodological issue of interest is that of the value of fully coupled ocean–atmosphere prediction systems (e.g., Stockdale et al. 1998) as opposed to two-tiered prediction systems.

One of the ultimate goals of the IRI is to produce the most skillful and reliable climate forecasts that are relevant to society's needs. Progress toward this goal involves automating the forecast process where possible without degrading forecast quality, to free time for researching and implementing further improvements. A step in this direction has been described here.

ACKNOWLEDGMENTS. The authors are grateful for the early comments of Chester Ropelewski, as well as the insightful comments and recommendations of the anonymous reviews.

REFERENCES

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., and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate,* **9**, 2660–2697.

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

CSIRO, cited 2002: Report of the ARGO Science Team 4th Meeting (AST-4), March 12–14, 2002, CSIRO Division of Marine Sciences, Hobart, Tasmania, Australia. [Available online at http://www.argo.ucsd.edu/iast4.pdf.]

Dai, A., K. E. Trenberth, and T. R. Karl, 1998: Global variations in droughts and wet spells: 1900–1995. *Geophys. Res. Lett.,* **25**, 3367–3370.

Deutsches Klimarechenzentrum, 1992: The ECHAM-3 Atmospheric General Circulation Model. Tech. Rep. 6, 189 pp. [Available from the Modellbetreuungsgruppe, Deutsches desstr. 55, Hamburg D-20146, Germany.]

DeWitt, D. G., 1996: The effect of cumulus convection on the climate of the COLA general circulation model. COLA Tech. Rep. 27, 58 pp. [Available from COLA, 4041 Powder Mill Road, Suite 302, Calverton, MD, 20705–3106.]

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

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate,* **12**, 1974–1989.

Gates, W. L., 1992: AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.,* **73**, 1962–1970.

——, and Coauthors, 1999: An overview of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.,* **80**, 29–55.

Goddard, L., and S. J. Mason, 2002: Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Climate Dyn.,* **19**, 619–632, doi:10.1007/s00382-002-0251-y.

——, A. G. Barnston, and S. J. Mason, 2003a: Evaluation of the IRI's "Net Assessment" Seasonal Climate Forecasts: 1997–2001. *Bull. Amer. Meteor. Soc.,* **84**, 1761–1781.

——, S. J. Mason, and A. W. Robertson, 2003b: Multimodel ensembling: Combining and refining. *Abstract, EGS–AGU–EUG Joint Assembly, Session on Intraseasonal and Seasonal Climate Predictability,* Nice, France, Amer. Geophys. Union, CD-ROM.

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

Graham, N. E., J. Michaelsen, and T. P. Barnett, 1987: An investigation of the El Niño–Southern Oscillation cycle with statistical models. Part 2: Model results. *J. Geophys. Res.,* **92**, 14 271–14 289.

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.

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.

Huberty, C. J., 1994: *Applied Discriminant Analysis.* Wiley, 466 pp.

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 seasonal forecast system 2000. *Bull. Amer. Meteor. Soc.,* **83**, 1019–1037.

Karl, T. R., R. W. Knight, and N. Plummer, 1995: Trends in high-frequency climate variability in the 20th century. *Nature,* **277**, 211–220.

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.

Kinter, J. L., III, J. Shukla, L. Marx, and E. K. Shneider, 1988: A simulation of winter and summer circulations with the NMC global spectral model. *J. Atmos. Sci.,* **45**, 2486–2522.

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

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

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

Mason, S. J., 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.

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

——, N. E. Graham, J. S. Galpin, and L. Goddard, 1999b: Estimating the "correct" probabilities of climate events. *Proc. 24th Annual Climate Diagnostics and Prediction Workshop,* Tucson, AZ, U.S. Department of Commerce, 210–213.

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

New, M., M. Hulme, and P. D. Jones, 1999: Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate,* **12**, 829–856.

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

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.,* **127**, 279–304.

——, C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126**, 2013–2033.

Pavan, V., and J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over the Euro-Atlantic: Skill scores and dynamic features. *Climate Dyn.,* **16**, 611–625.

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, H. M. van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.,* **107**, 4710, doi:10.1029/2002JD002712.

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.

Roeckner, E., and Coauthors, 1992: Simulation of the present-day climate with the ECHAM model: Impact of model physics and resolution. Max-Planck-Institut für Meteorologie 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-Institut für Meteorologie Rep. 218, Hamburg, Germany, 90 pp.

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.

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., A. J. Busalacchi, M. J. McPhaden, A. D. Moura, G. Reverdin, M. Vianna, and S. E. Zebiak, 1998: A pilot research moored array in the tropical Atlantic (PIRATA). *Bull. Amer. Meteor. Soc.,* **79**, 2019–2031.

Stockdale, T. N., D. L. T. Anderson, J. O. S. Alves, and M. A. Balmaseda, 1998: Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature,* **392**, 370–373.

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

Toth, Z., and S. Vannitsem, 2002: Model errors and ensemble forecasting. *Proc. Eighth ECMWF Workshop on Meteorological Operational Systems,* Reading, United Kingdom, ECMWF, 146–154.

van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for near normal often fail? *Wea. Forecasting,* **6,** 76–85.

Wallace, J. M., C. Smith, and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J. Climate,* **5**, 561–576.

Ward, M. N., and A. Navarra, 1997: Pattern analysis of SST-forced variability in ensemble GCM simulations: Examples over Europe and the tropical Pacific. *J. Climate,* **10**, 2210–2220.

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

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

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

Wilson, S., 2000: Launching the Argo armada. *Oceanus,* **42**, 17–19.

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

Zhou, J. L., and A. L. Tits, 1993: Nonmonotone line search for minimax problems. *J. Optim. Theory Appl.,* **76**, 455–476.