NOTES AND CORRESPONDENCE

Interpretation of Seasonal Climate Forecast Using Brier Skill Score, The Florida State University Superensemble, and the AMIP-I Dataset

L. STEFANOVA AND T. N. KRISHNAMURTI

Department of Meteorology, The Florida State University, Tallahassee, Florida

3 July 2000 and 28 September 2001

ABSTRACT

The superensemble technique has been proven to be successful in producing a deterministic forecast superior not only to any of the individual models going into it, but also to the multimodel ensemble forecast. Research so far has been done on the superensemble as a deterministic forecast, and it has been shown that using the superensemble method leads to a significant reduction in rms errors. This paper investigates the skill of the superensemble as a *probabilistic forecast,* and it compares it with that of the multimodel ensemble. Using the Atmospheric Model Intercomparison Project (AMIP I) seasonal multimodel precipitation forecasts, probability forecasts are defined for the multimodel ensemble and for the multimodel superensemble. The Brier skill score of these forecasts is calculated for different thresholds of precipitation anomaly. It is shown that both the multimodel ensemble and the superensemble probability forecasts are much better than climatological forecast and that the superensemble probability forecast outperforms the multimodel bias-removed ensemble at any threshold level.

1. Introduction

Various methods of ensemble averaging have been carried out in previous studies by numerous authors. Examples include Thompson (1977), Fraedrich and Smith (1989), Wobus and Kalnay (1995), Sarda et al. (1996), and Pavan and Doblas-Reyes (2000). The superensemble is a technique developed by Krishnamurti et al. (1999) that produces a single forecast derived from a multimodel set of forecasts. It differs from a conventional bias-removed multimodel ensemble in that the different models are weighed by sets of regression coefficients obtained during a training period previous to the superensemble forecast mode. These regression coefficients associated with each individual model conceivably can be interpreted as a measure of that model's relative reliability for the given point over the training period. That this is a reasonable interpretation is suggested by the successful comparisons of probability forecasts using this assumption versus a conventional probability forecast generated from a conventional multimodel bias-removed ensemble. Further substantiation of the regression coefficient interpretation is offered at the end of the following section. A multimodel bias-removed ensemble is equivalent to the special case of a superensemble of models that have the same history of reliability and no scaling bias. Analysis of the rms errors and correlations has demonstrated that the superensemble forecast is superior not only to all the models going into constructing it, but also to the multimodel bias-removed ensemble.

Because the underlying assumption is that not all models are equally reliable in different points in space, it is to be expected that the probability associated with the resulting deterministic forecast would not be the same for the superensemble and the multimodel ensemble. The probability of a forecast event from an ensemble system is based on the fraction of ensemble members predicting that event. After suitably defining the corresponding probability using the superensemble method, it is desirable to compare the relative qualities of the two probabilistic forecasts.

A standard measure of the skill of probabilistic forecasts, similar to the rms errors used for assessing the skill of deterministic forecasts, is the Brier skill score. First developed by Brier (1950), it has become widely used in association with ensemble forecasts from a single model and with multimodel ensemble (Murphy and Katz 1985; Sperber and Palmer 1996). The Brier skill score measures the improvement of probabilistic forecast relative to forecast of climatological probability.

The Atmospheric Model Intercomparison Project da-

Corresponding author address: T. N. Krishnamurti, Dept. of Meteorology, The Florida State University, 404 Love Building, Tallahassee, FL 32306-4520.
E-mail: tnk@io.met.fsu.edu

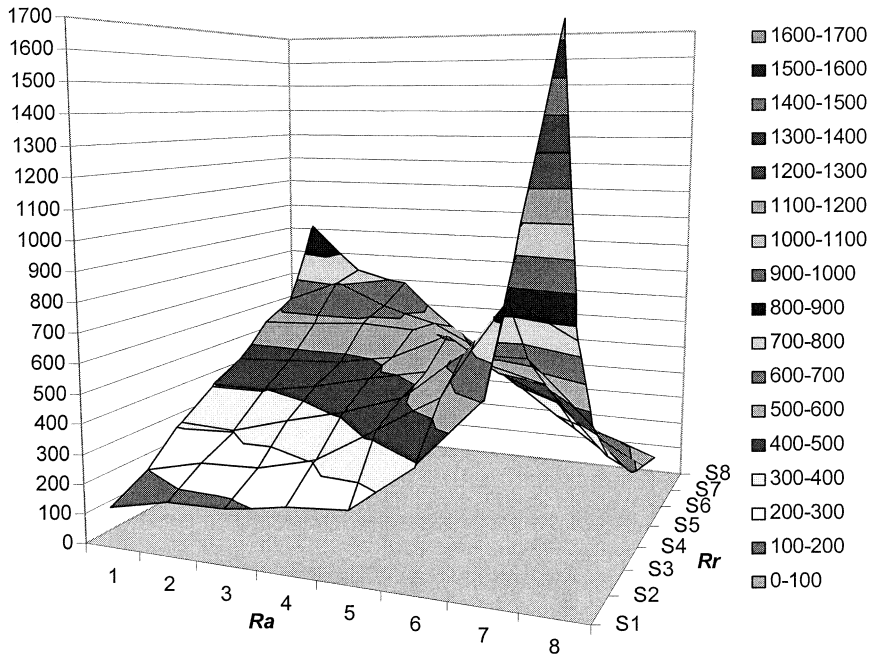

FIG. 1. Frequency of occurrence of all $(R_a \mid R_r)$ combinations throughout the dataset.

taset (AMIP I) is described in detail by Gates (1992). The data used here are decade-long series of monthly mean global precipitation forecasts from a subset of eight of the AMIP-I models: those of the Bureau of Meteorology Research Centre, Commonwealth Scientific and Industrial Research Organisation, European Centre for Medium-Range Weather Forecasts, Geophysical Fluid Dynamics Laboratory, Laboratoire de Météorologie Dynamique, Max-Planck-Institut für Meteorologie, National Meteorological Center (now known as the National Centers for Environmental Prediction), and the Met Office. Subsets of two adjacent years are sequentially removed from the dataset, and the remaining eight years are used as a training period for generating the model mean, the observed mean, the regression coefficients, and the statistical weights used for the superensemble probability forecast, as described in section 4 of this paper. The two years left out of the corresponding training period are used for testing the probabilistic forecast skill of the superensemble and the conventional multimodel bias-removed ensemble.

2. Superensemble

The superensemble forecast is constructed as

$$S = \sum_{i=1}^{N} a_i(F_i - \overline{F_i}) + \overline{O}, \tag{1}$$

where F_i is the ith model forecast, $\overline{F_i}$ is the mean of the ith forecast over the training period, $\overline{O}$ is the observed mean over the training period, a_i are regression coefficients obtained by a minimization procedure during the training period, and N is the number of forecast models involved (8 in our case).

The coefficients a_i are derived from estimating the minimum of the function

$$G = \sum_{t=1}^{T_{\text{train}}} (S_t - O_t)^2, \tag{2}$$

that is, by minimizing the mean square error of the forecast. The a_i vary in space but are constant in time.

When (1) is compared to a multimodel bias-removed ensemble,

$$E = \frac{1}{N} \sum_{i=1}^{N} (F_i - \overline{F_i}) + \overline{O}, \tag{3}$$

it is obvious that, in addition to removing bias, the superensemble scales the individual model forecast's contributions according to their relative performance during the training period in a way that, mathematically, is equivalent to weighting them.

To demonstrate that the regression coefficients are an estimate of the relative reliability of a forecast model (in terms of rms error, and in the context of the AMIP-I simulation), the following calculation is done. For each point in the dataset, the regression coefficients for each of the eight models are ranked according to their relative size. The smallest coefficient is ranked $R_a = 1$, the next-to-smallest one is ranked $R_a = 2$, and so on, and the largest coefficient is ranked $R_a = 8$. The rms errors are similarly ranked from $R_r = 1$ to $R_r = 8$, going from smallest to largest error. The frequency of occurrence of each combination $(R_a \mid R_r)$ is shown in Fig. 1. The resulting plot has a saddle shape, with the highest values

at ($R_a = 1 \mid R_r = 8$) and ($R_a = 8 \mid R_r = 1$). In other words, the smallest regression coefficients tend to correspond to the largest rms errors, and the largest regression coefficients tend to correspond to the smallest rms errors. The least frequent combinations are ($R_a = 1 \mid R_r = 1$) and ($R_a = 8 \mid R_r = 8$). The crest of the frequency surface is almost exactly at the diagonal [($R_a = m \mid R_r = 9 - m$), $m = 1, 8$]. This provides an empirical backing to the notion that the regression coefficients can be considered statistically as estimates of a model's relative reliability and justifies their being regarded as weights in (1).

3. Probabilistic forecasting

A probabilistic forecast is one that estimates the probability of occurrence of a chosen event ε. The event type selected for this study is ''precipitation rate anomaly relative to the mean state exceeding a preselected threshold level.''

For an ensemble of *equally reliable* models the probability P of the event ε is $(m/N) \times 100\%$, where m is the number of ensemble members forecasting ε and N is the total number of ensemble forecasts. The issue of defining the probability of an event based on the superensemble is somewhat more complicated and is discussed in a separate section.

Because for a single realization a probability forecast is neither correct nor wrong, probability forecasts are verified by analyzing the joint (statistical) distribution of forecasts and observations. One of the most widely used methods for verification of probability forecasts is the Brier skill score.

To define the Brier skill score, select an event ε that either happens at realization k or does not [$o(k) = 1$ if ε occurred, $o(k) = 0$ if it did not] and is forecast to occur with probability $f(k)$. Following Wilks (1995), the Brier score is then defined as

$$b = \frac{1}{n}\sum_{k=1}^{n}[f(k) - o(k)]^2,$$

where the index k refers to the forecast–observation pairs and n is the total number of such pairs within the dataset. The lowest possible value of the Brier score is obviously zero, and it can only be achieved with a perfect deterministic forecast.

Let the probabilistic forecast for ε be done within I discrete categories y_i. The frequency with which forecasts of y_i are issued is $p(y_i)$. The frequency within a category y_i forecast with which the event ε actually occurs is the conditional frequency $\overline{o_i} = p[o(k) = 1 \mid y_i]$. A reliability diagram is a plot of $\overline{o_i}$ versus y_i, accompanied by the forecast frequency distribution $p(y_i)$ versus y_i. For a perfect forecast, the reliability diagram would be a line at 45°.

As suggested by Murphy (1973), it is useful to decompose the Brier score into three terms: reliability, resolution, and uncertainty, as follows.

$$b = \underbrace{\sum_{i=1}^{I} p(y_i)(y_i - \overline{o_i})^2}_{\text{reliability}} - \underbrace{\sum_{i=1}^{I} p(y_i)(\overline{o_i} - \overline{o})^2}_{\text{resolution}} + \underbrace{\overline{o}(1 - \overline{o})}_{\text{uncertainty}}$$
$$= b_{\text{rel}} - b_{\text{res}} + b_{\text{unc}},$$

where $\overline{o} = (1/n)\Sigma_{k=1}^{n} o(k)$ is the unconditional mean frequency of occurrence of the event ε. The reliability term evaluates the statistical accuracy of the forecast—a perfectly reliable forecast is one for which the observed conditional frequency $\overline{o_i}$ is equal to the forecast probability (i.e., over all forecasts for y-percent chance of ε, ε will occur in y percent of the times).

The resolution term addresses the distance between the forecast frequency and the unconditional climatological frequency. Forecasts that are always close to the climatological frequency exhibit good reliability, because the forecast frequency matches the observed frequency, but poor resolution, because they are not able to distinguish between different regimes.

The uncertainty term is a measure of the variability of the system and is not influenced by the forecast.

Skill scores are calculated with respect to a reference forecast as $B = (b - b_{\text{ref}})/(b_{\text{perf}} - b_{\text{ref}}) = 1 - (b/b_{\text{ref}})$, given that b_{perf} is 0. If a climatological forecast ($b_{\text{res}} = 0$, $b_{\text{rel}} = 0$) is taken as a reference, $B = 1 - (b/b_{\text{unc}})$, $B_{\text{rel}} = 1 - (b_{\text{rel}}/b_{\text{unc}})$, and $B_{\text{res}} = b_{\text{res}}/b_{\text{unc}}$. For a *perfect* forecast system, $B = B_{\text{rel}} = B_{\text{res}} = 1$. For a *climatological* forecast, $B = B_{\text{rel}} = B_{\text{res}} = 0$.

4. Assigning probability to a superensemble forecast

The superensemble defined by (1) can be rewritten in a form that resembles an ensemble of modified unbiased forecasts Fs_i:

$$S = \left[\sum_{i=1}^{N} a_i(F_i - \overline{F_i})\right] + \overline{O} = \frac{1}{N}\sum_{i=1}^{N}[Na_i(F_i - \overline{F_i}) + \overline{O}]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\text{Fs}_i,$$

where

$$\text{Fs}_i = Na_i(F_i - \overline{F_i}) + \overline{O} \tag{4}$$

are the modified forecasts. One possibility for defining the probability of the superensemble then is to treat it as an ensemble of the forecasts Fs_i. Such an approach, however, would discount the fact that the forecasts are not equally reliable. If the ith regression coefficient is very small, the corresponding model's contribution to the superensemble would be relatively insignificant. It is then necessary for the purpose of probability forecast to attach additional weights to the modified forecasts given by (4).

The probability of ε derived from the superensemble can be defined as

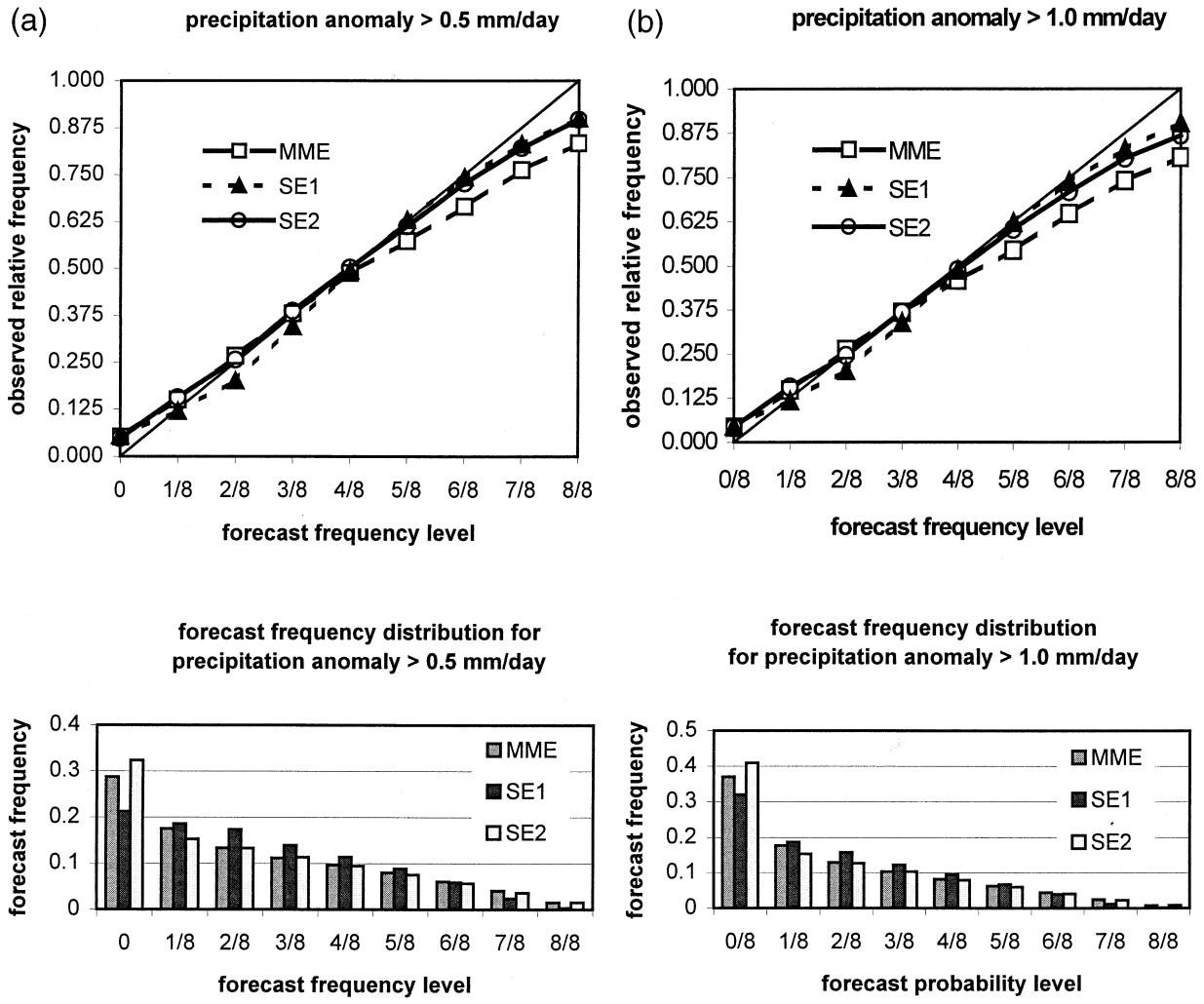

FIG. 2. Reliability diagrams for precipitation anomaly exceeding, respectively, (a) 0.5 ($\bar{o} = 0.29$), (b) 1.0 ($\bar{o} = 0.24$), (c) 1.5 ($\bar{o} = 0.19$), and (d) 2.0 mm day^{-1} ($\bar{o} = 0.16$) on a monthly mean scale using cross validation. The long-dashed line with open squares is from MME, the short-dashed line with filled triangles is from SE1, and the solid line with open circles is from SE2. Frequency distributions are on the plots beneath with triplets of bars corresponding to MME, SE1, SE2, from left to right.

$$P_\varepsilon = \sum_{i=1}^{N} w_i \delta(\mathrm{Ps}_i),$$

where

$$\delta(\mathrm{Ps}_i) = \begin{cases} 0 & \mathrm{Fs}_i \notin \varepsilon \\ 1 & \mathrm{Fs}_i \in \varepsilon, \end{cases}$$

where the weights w_i are normalized so that their sum is unity. For equally reliable models, $w_i = 1/N$.

One way of defining the weights associated with the different models making up the superensemble is to relate them to the absolute values of the corresponding regression coefficients, that is, as

$$w_i = \frac{|a_i|^\lambda}{\sum_{i=1}^{N} |a_i|^\lambda}.$$

The best choice for λ empirically is 0.5. This method will be referred to as method 1, or SE1.

An alternative method that will be shown to yield better results is based on relating the statistical weights to the hit rate for the event and nonevent during the training period:

$$w_i = \frac{c_i^\kappa}{\sum_{i=1}^{N} c_i^\kappa},$$

where c_i is the sum of the hit rate for the event and the hit rate for the nonevent for the ith model over the training period and κ is an empirically chosen constant fixed, as best choice, at 3. To obtain these, a contingency table is calculated for the event E for each modified model Fs_i over the training period such that α is the number of events forecast and observed, β is the number

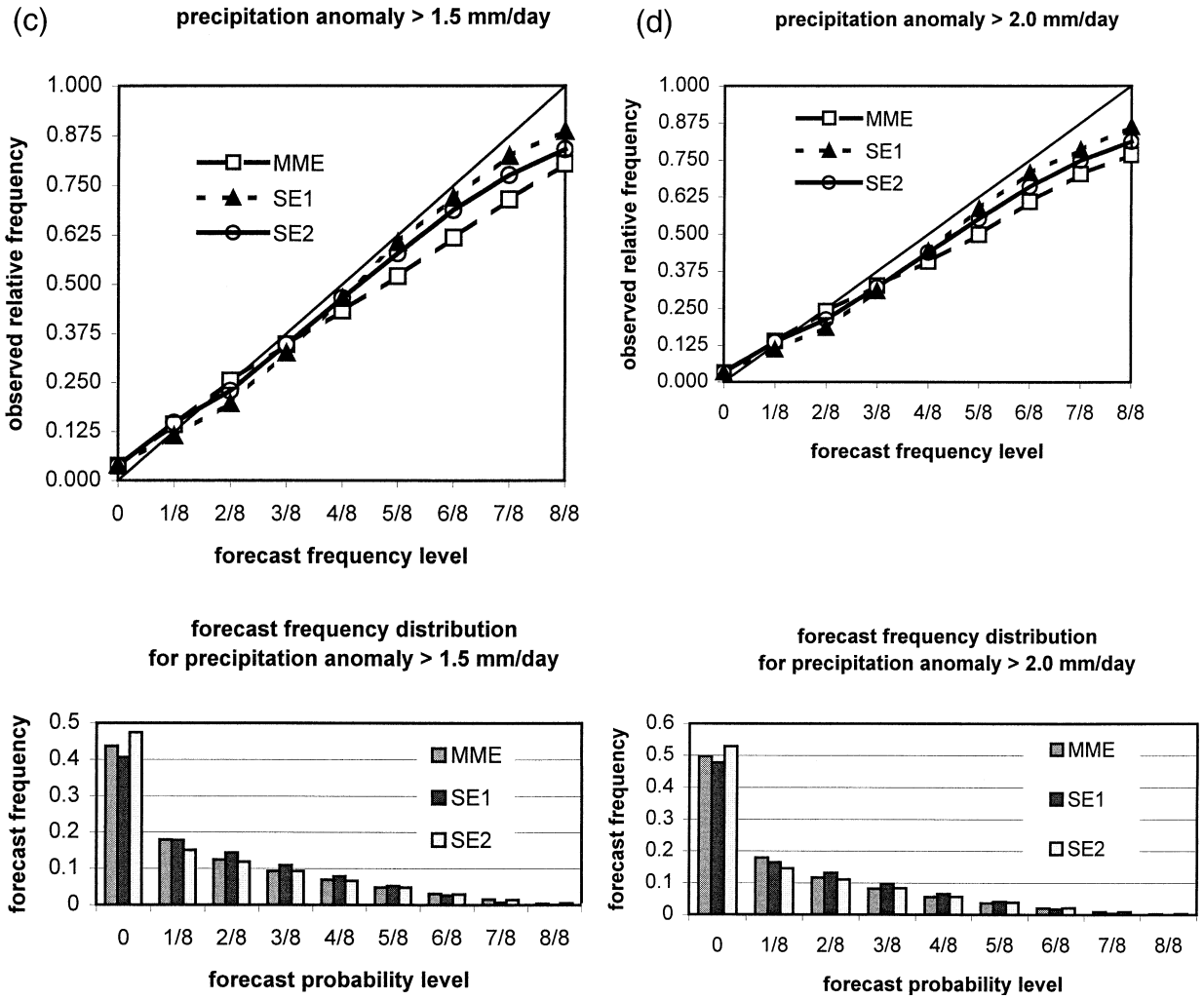

FIG. 2. (*Continued*)

of events forecast but not observed, γ is the number of events observed but not forecast, and δ is the number of events neither forecast nor observed. Then,

$$c_i = \frac{\alpha}{\alpha + \gamma} + \frac{\delta}{\beta + \delta}.$$

This method of defining the weights is event dependent and can better represent the relative trustworthiness of the different models with respect to the chosen event. Note that with either definition the weights are varying in space; that is, the different models have varying relative contributions to the total probability depending on the spatial location of the point in question. This method is further referred to as method 2, or SE2.

Both SE1 and SE2 probability forecasts are compared with the conventionally defined probability forecasts from the multimodel bias-removed ensemble [see (3)], where all N individual unbiased forecasts $F_i - \overline{F} + \overline{O}$ are weighted equally with a weight of $1/N$. This method is denoted as MME.

5. Results

Reliability diagrams are shown in Figs. 2a–d. Results are shown for MME and for the superensemble calculated by both methods discussed above. The events are precipitation anomalies with respect to the series mean exceeding a threshold (0.5, 1, 1.5, 2) mm day^{-1} in the monthly mean for all points of the global Tropics from 35°S to 35°N using cross validation. The abscissa is the forecast probability level y_i, and the ordinate is the observed conditional frequency $\overline{o_i}$. The frequency of use of forecasts of probability y_i for each forecasting system (MME, SE1, and SE2) is shown underneath each plot. The unconditional frequency of occurrence of the event over the entire sample is indicated in the figure caption. The Brier skill score values for reliability and resolution (B_{rel} and B_{res}), along with the total score B are shown in Table 1. For all threshold levels the skill scores are positive, which means they offer improvement over the climatological forecast.

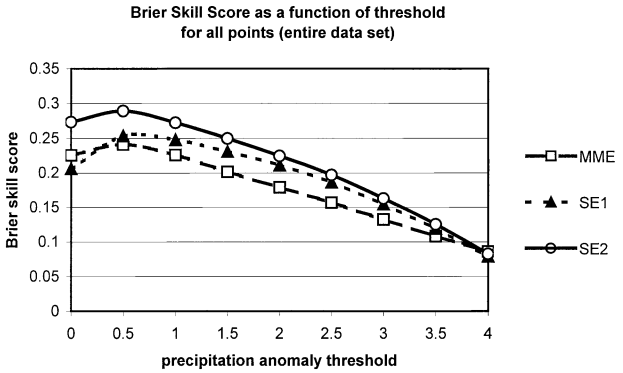

FIG. 3. Total cross-validated Brier skill score for MME (long-dashed line with open squares), SE1 (short-dashed line with filled triangles) and SE2 (solid line with open circles) as a function of the threshold level (mm day^{-1}) of the monthly mean precipitation anomaly.

Both the multimodel ensemble and the superensemble have a very strong reliability (B_{rel} is very close to 1). Although the reliability skill scores are similar, the superensemble forecast is visually much closer to the 45° line, implying higher reliability. The reason this effect is not strongly reflected in the reliability skill score lies in the relatively low frequency of use of high-probability forecasts; thus, the contribution of these points to the overall skill is low. Although the skill score does not benefit significantly from this improvement, it is an important practical conclusion that high-probability forecasts by the SE1 or SE2 are, statistically, more trustworthy then those by MME. In terms of total Brier skill score, the majority of the improvement of SE over MME comes from the resolution term.

The total skill score for MME, SE1, and SE2 as a function of the precipitation anomaly threshold is shown in Figs 3 and 4. The superensemble calculated by either method produces a probability forecast that has roughly 25% improvement over MME.

As discussed by Palmer et al. (2000), a Brier skill score of 1, which is what a perfect deterministic forecast would score, is not a reasonable expectation for a probability forecast. If the forecast were constructed using ensemble members from the same model with perturbed initial conditions, then an estimate of a reasonable upper limit for Brier skills score could be made assuming that the ensemble spread reflects the internal variability of the system. However, given that in this case the ensemble members come from different models it would not be adequate to consider the variability between them as reflecting the variability of the system, and thus no estimate of the upper limit on Brier skill score can be calculated.

Because the Brier skill score was defined initially for a time series at a single point, part of its interpretation becomes dubitable when different points in space are grouped into one time series. Because the sample climatological mean of the frequency of any event varies

TABLE 1. Brier skill scores for reliability (B_{rel}), resolution (B_{res}), and total (B) for different levels of precipitation anomaly (PA) threshold (mm day^{-1}).

	PA > 0.0			PA > 0.5			PA > 1.0			PA > 1.5			PA > 2.0			PA > 2.5		
	B_{rel}	B_{res}	B	B_{rel}	B_{res}	B	B_{rel}	B_{res}	B	B_{rel}	B_{res}	B	B_{rel}	B_{res}	B	B_{rel}	B_{res}	B
MME	0.99	0.24	0.23	0.99	0.25	0.24	0.99	0.24	0.23	0.98	0.22	0.20	0.98	0.20	0.18	0.98	0.18	0.16
SE1	0.97	0.22	0.21	0.99	0.26	0.25	0.99	0.25	0.25	0.99	0.24	0.23	0.99	0.23	0.21	0.98	0.21	0.19
SE2	0.99	0.28	0.27	0.99	0.30	0.29	0.99	0.28	0.27	0.99	0.26	0.25	0.99	0.24	0.22	0.98	0.22	0.20

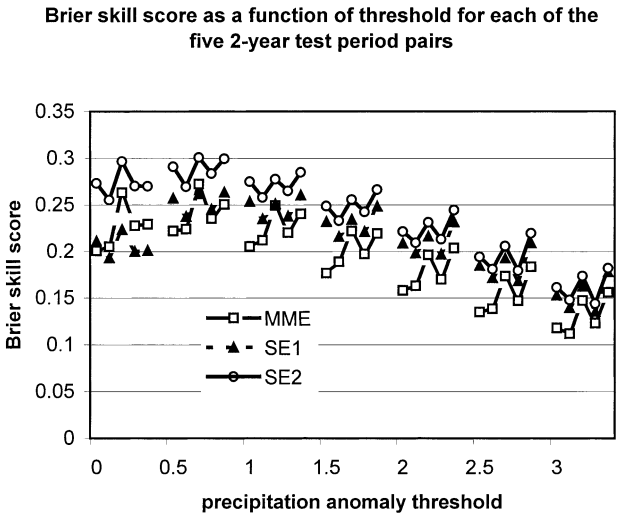

FIG. 4. As in Fig. 3, except that the Brier skill scores for all five 24-month-long test periods are shown separately in five-point sets corresponding to the given threshold level.

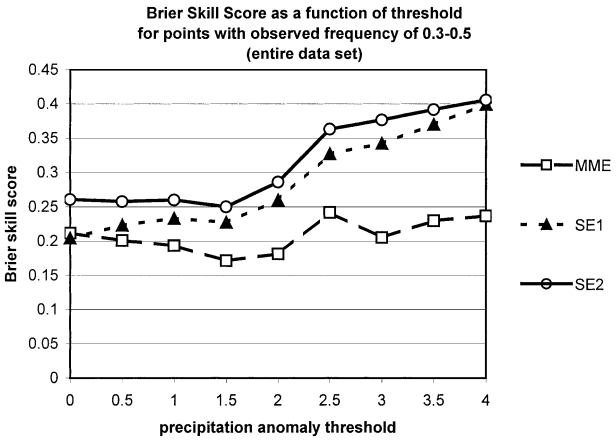

FIG. 5. As in Fig. 3, but only for points with observed frequency 0.3–0.5 of precipitation anomaly above threshold value.

among geographical locations, the meaning of the ''resolution'' term can be washed out, because it describes the distance from the overall (averaged in time *and* space) sample climatological mean. Thus, it is possible that the Brier skill score attributes undeserved skill to forecasts that are far away from the overall mean frequency, although they may be at the corresponding point's climatological frequency, or conversely attributes undeserved penalty for forecasts far from the point's climatological frequency but close to the overall mean frequency. To discount for such effects, calculations of Brier skill score were repeated over points with similar climatological frequency of the given event. The Brier skill score as a function of threshold for all points whose climatological frequency of the event is between 0.3 and 0.5 is shown in Fig. 5. The conclusions made on the basis of the entire global Tropics, including points with a variety of climatological frequencies of events, hold true even more strongly—the superensemble is providing a better probability forecast than is the multimodel ensemble for all threshold levels.

6. Concluding remarks

AMIP-I datasets come from decade-long integration of several atmospheric general circulation modeling groups. These datasets all include prescribed sea surface temperatures and sea ice. In Krishnamurti et al. (1999, 2000), the results obtained from a multimodel superensemble based on several of these multimodels were examined using standard anomaly correlations and root-mean-square errors as means of skill for the evaluation of seasonal and multiseasonal forecasts, based on these models and the superensemble.

In these earlier studies, it was noted that the superensemble forecasts do have a higher skill when compared with the member model forecasts, the ensemble mean forecasts, and the ensemble mean of bias-removed individual models. The aforementioned skills were based on a deterministic forecast view of the multimodels and the superensemble. It was, however, felt that a more stringent test was necessary to show the forecast skill of the superensemble, in which a probabilistic view could be invoked. The Brier skill score is a very conservative test of the performance of models in which the probabilities' measures are examined. In this paper, we use different thresholds of monthly predicted precipitation by the different models and the superensemble. It is possible to demonstrate clearly that the probabilistic forecasts of the superensemble are indeed better than the climatological forecasts and the multimodel ensemble forecasts. After an examination of the reliability component of the Brier skill score at several thresholds, it is concluded that this contributes to improved probability forecasts as compared with the multimodels and climatological forecasts. It is also noted that a large proportion of the probability forecast skill of the superensemble over other forecast representations comes from the resolution component of the Brier skill score. Overall, it is shown that the superensemble provides a better probability forecast when compared with the multimodel ensemble and the climatological forecast at all precipitation thresholds. The future applications of the multimodel superensemble are expected to migrate toward coupled atmosphere–ocean member models. Those models, for which the seasonal forecast skills are assessed from the Brier skill scores, would help us to determine their usefulness above those of a climatological forecast.

Acknowledgments. This work was supported by NOAA Grants NA86GP0031, NA96GPO400, NA76GPO521, and NA77WA0571 and Florida State University Research Foundation Grant 1368-760-45 to the FSU Climate Center. We wish to acknowledge the

Lawrence Livermore Laboratory in California for providing the AMIP dataset used here.

REFERENCES

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

Fraedrich, K., and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting. *J. Climate,* **2,** 291–294.

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

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, E. Williford, S. Gadhil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel superensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

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

——, and R. W. Katz, 1985: *Probability, Statistics and Decision Making in the Atmospheric Sciences.* Westview, 545 pp.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multimodel ensemble integration. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2033.

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

Sarda, J., G. Plant, C. Pires, and R. Vantard, 1996: Statistical and dynamical long range atmospheric forecasts: Experimental comparison and hybridization. *Tellus,* **48A,** 518–537.

Sperber, K. R., and T. N. Palmer, 1996: Interannual tropical rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project. *J. Climate,* **9,** 2727–2750.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.

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

Wobus, R. L., and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC. *Mon. Wea. Rev.,* **123,** 2132–2148.