NOTES AND CORRESPONDENCE

Two Extra Components in the Brier Score Decomposition

D. B. STEPHENSON

School of Engineering, Computing, and Mathematics, University of Exeter, Exeter, United Kingdom

C. A. S. COELHO

Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais, Cachoeira Paulista, São Paulo, Brazil

I. T. JOLLIFFE

School of Engineering, Computing, and Mathematics, University of Exeter, Exeter, United Kingdom

(Manuscript received 20 December 2006, in final form 20 July 2007)

ABSTRACT

The Brier score is widely used for the verification of probability forecasts. It also forms the basis of other frequently used probability scores such as the rank probability score. By conditioning (stratifying) on the issued forecast probabilities, the Brier score can be decomposed into the sum of three components: uncertainty, reliability, and resolution. This Brier score decomposition can provide useful information to the forecast provider about how the forecasts can be improved.

Rather than stratify on all values of issued probability, it is common practice to calculate the Brier score components by first partitioning the issued probabilities into a small set of bins. This note shows that for such a procedure, an additional two within-bin components are needed in addition to the three traditional components of the Brier score. The two new components can be combined with the resolution component to make a generalized resolution component that is less sensitive to choice of bin width than is the traditional resolution component. The difference between the generalized resolution term and the conventional resolution term also quantifies how forecast skill is degraded when issuing categorized probabilities to users. The ideas are illustrated using an example of multimodel ensemble seasonal forecasts of equatorial sea surface temperatures.

1. Introduction

The Brier score is the mean squared difference between issued forecast probabilities and observed binary outcomes (Brier 1950; Jolliffe and Stephenson 2003). It is one of the oldest and most commonly used scores for assessing the skill of probability forecasts of binary events (e.g., rain or no rain) and it forms the basis of other widely used probability scores such as the ranked probability score (Epstein 1969).

It can be revealing to decompose the Brier and ranked probability scores into the sum of three components: forecast *reliability* (bias of conditional means), forecast resolution (variance of conditional means), and observational uncertainty (Sanders 1963; Murphy 1971, 1973, 1986). To do this, it is necessary to calculate the mean of the observations (the relative frequency of the observed event) stratified/conditioned on the different forecast probabilities. One can do this either by dividing the data into a finite set of categories (bins) of forecast probability *or* by directly stratifying on each of the distinct probability values that have been issued. The early studies such as Murphy (1971) and others used the latter unbinned approach and stratified directly on each of the issued probability values (e.g., $f = 0.1, 0.2, \ldots, 0.9$ for subjective probability forecasts). The Brier score decomposition assumed such direct stratification.

Corresponding author address: Dr. David B. Stephenson, School of Engineering, Computing, and Mathematics, University of Exeter, North Park Rd., Exeter EX4 4QF, United Kingdom.
E-mail: d.b.stephenson@exeter.ac.uk

DOI: 10.1175/2007WAF2006116.1

However, it is now common practice to calculate the components by stratifying over bins of probabilities such as those used to produce reliability diagrams (e.g., Atger 2003). This widespread usage has even led to the misconception that "estimating reliability and resolution requires a categorization of probabilistic forecasts" (Atger 2004, p. 628). Nevertheless, the effects of binning need to be considered since categorized probability forecasts are often what are finally issued to the forecast user. For example, the recent System 3 seasonal forecasting system at the European Centre for Medium-Range Weather Forecasts (ECMWF) issues probabilities for temperature terciles in seven distinct unequal-width categories: 0–0.1, 0.1–0.2, 0.2–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7, and 0.7–1.0 (Dr. F. J. Doblas-Reyes 2007, personal communication). The reliability and resolution calculated from these binned probabilities are not guaranteed to be the same as the reliability and resolution components calculated using the uncategorized probabilities produced by the forecasting system.

In addition to simplifying the probability forecasts for users, binning can have several advantages. First, it acts as a crude form of smoothing thereby making the conditional means less uncertain and the reliability curve less noisy (Atger 2003). Second, larger bins can avoid sparseness problems that can occur when probabilities are rarely or never issued within smaller bins, for example, for a small sample of probability forecasts from a large ensemble system (Atger 2004). Third, it can allow cleaner comparison of Brier score components for forecasting systems having different numbers of ensemble forecasts (Mullen and Buizza 2002; Ferro 2007).

This study has mathematically investigated the decomposition of the Brier score in these more general situations where observations are stratified into bins of forecast probabilities rather than directly on the issued probability values. Section 2 of this paper shows that the Brier score is no longer identical to a sum of just three components but also has two additional components that account for within-bin variations. Section 3 illustrates this with a multimodel ensemble forecasting example. Section 4 presents some concluding remarks and possible ideas for future work.

2. The Brier score decomposition

a. Basic definitions

This section will introduce the mathematical definitions needed to calculate the decomposition of the Brier score defined for a historical sample of paired binary observations (o) and probability forecasts (f).

First, consider a partition of the probability unit interval [0, 1] into m mutually exclusive subintervals (bins) labeled by the index $k = 1, 2, \ldots, m$. Denote the n_k probability forecasts that have fallen in the kth bin by f_{kj} where $j = 1, 2, \ldots, n_k$. The total number of forecasts in all bins is $n = \Sigma_{k=1}^{m} n_k$. The overbar symbol will be used to denote averages within a particular bin; for example, the average probability of all the probability forecasts in the kth bin is given by $\bar{f}_k = (1/n_k)\Sigma_{j=1}^{n_k} f_{kj}$. Let the variable o_{kj} denote the binary outcome (0 or 1) of the observed event associated with the jth probability forecast in the kth bin (i.e., the one whose probability forecast is f_{kj}). Note that sufficiently narrow bins overlapping the issued probability values can always be chosen so that only one probability value (yet perhaps several forecasts) occurs within each bin.

The *forecast error* for the jth forecast in the kth bin is then given by $f_{kj} - o_{kj}$ and so the mean squared forecast error is

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj} - o_{kj})^2, \qquad (1)$$

which is known as the Brier score (Brier 1950). The Brier score is a *negatively oriented* score that gives smaller values for better forecasts.

b. Brier score decomposition

The Brier score can be rewritten as a nested mean of within-bin averages:

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{m} n_k\left[\frac{1}{n_k}\sum_{j=1}^{n_k}(f_{kj} - o_{kj})^2\right]. \qquad (2)$$

The expression inside the brackets is the mean of a squared quantity and so can be written as the sum of the square of the mean quantity plus the variance of the quantity. In other words, the within-bin mean of the squared forecast error is equal to the sum of the squared mean forecast error and the *within-bin variance* of the forecast error, $f_{kj} - o_{kj}$:

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{m} n_k\left[(\bar{f}_k - \bar{o}_k)^2 + \frac{1}{n_k}\sum_{j=1}^{n_k}(f_{kj} - o_{kj} - \bar{f}_k + \bar{o}_k)^2\right], \qquad (3)$$

where

$$\bar{o}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} o_{kj}$$

is the relative frequency that the observed event occurred at times when forecast probabilities were in the interval $[p_k, p_{k+1}]$. Expanding the square in the final summation of (3) gives two sums of squares terms (variances) and a cross-product (covariance) term, so that the Brier score becomes

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{m} n_k\left[(\bar{f}_k - \bar{o}_k)^2 + \frac{1}{n_k}\sum_{j=1}^{n_k}(o_{kj} - \bar{o}_k)^2 + \frac{1}{n_k}\sum_{j=1}^{n_k}(f_{kj} - \bar{f}_k)^2 - \frac{2}{n_k}\sum_{j=1}^{n_k}(o_{kj} - \bar{o}_k)(f_{kj} - \bar{f}_k)\right]. \tag{4}$$

The first two terms in this expression are the traditional components of the Brier score. The first term

$$\frac{1}{n}\sum_{k=1}^{m} n_k(\bar{f}_k - \bar{o}_k)^2$$

in (4) summarizes the unconditional and conditional bias in the forecasts and is known as the *reliability* of the forecasts (REL). In principle, it can be reduced by good calibration of the forecasts (Murphy 1986). The second term in (4) can be written as the total variance minus the variance of the within-bin means of the observed variable:

$$\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k} n_k(o_{kj} - \bar{o}_k)^2 = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k} n_k(o_{kj} - \bar{o})^2 - \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k - \bar{o})^2, \tag{5}$$

where

$$\bar{o} = \frac{1}{n}\sum_{k=1}^{m} n_k\bar{o}_k$$

is the climatological base rate (mean probability) for the event to occur. The first term on the right-hand side of (5) can be shown to be $\bar{o}(1 - \bar{o})$ by expanding the square and noting that $o_{kj}^2 = o_{kj}$, for binary variables. Hence, the second term in (4) is given by

$$\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(o_{kj} - \bar{o}_k)^2 = \bar{o}(1 - \bar{o}) - \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k - \bar{o})^2, \tag{6}$$

that is, the observational *uncertainty* $\bar{o}(1 - \bar{o})$ (UNC) minus the forecast *resolution* (RES). Therefore, the first two terms in (4) give the traditional REL − RES + UNC components of the Brier score.

The third and fourth terms in (4) are pooled averages of the *within-bin variance* (WBV) of the forecasts minus the *within-bin covariance* (WBC) between forecasts and observations. Both terms vanish if only one value of probability is forecast for each bin ($f_{kj} = \bar{f}_k$ for all j). In other words, when there is no within-bin variation among the forecast probabilities (e.g., when stratifying on unbinned forecast values), then only the first two terms in (4) need to be considered in the decomposition of the Brier score; the Brier score is then simply equal to the well-known decomposition REL − RES + UNC. However, whenever there is any variation in forecast probabilities within any of the bins, then the Brier score becomes

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{f}_k - \bar{o}_k)^2 - \frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1 - \bar{o}) + \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj} - \bar{f}_k)^2 - \frac{2}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(o_{kj} - \bar{o}_k)(f_{kj} - \bar{f}_k). \tag{7}$$

The Brier score has five rather than three components: BS = REL − RES + UNC + WBV − WBC.

c. Generalized resolution

The two within-bin terms help compensate the decreasing resolution component when the bin size is increased. The three-component decomposition of the Brier score can be maintained by generalizing the resolution term to include the two within-bin terms:

$$\frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k - \bar{o})^2 - \frac{1}{n}\sum_{k=1}^{m} n_k\left[\frac{1}{n_k}\sum_{j=1}^{n_k}(f_{kj} - \bar{f}_k)^2 - \frac{2}{n_k}\sum_{j=1}^{n_k}(o_{kj} - \bar{o}_k)(f_{kj} - \bar{f}_k)\right].$$

In other words, by using the *generalized resolution* defined by GRES = RES − WBV + WBC, the Brier score becomes REL − GRES + UNC. As illustrated in the following section, the resulting generalized resolution component is less sensitive to the choice of bin size than is the classic definition of resolution.

We prefer to include these terms in the resolution term since unlike the reliability term, the within-bin terms cannot be easily transformed to zero by one-to-one recalibration of the probability forecasts. However, the within-bin terms can be made to vanish by mapping the forecast probabilities to a smaller number of bin location values as is often done when simplifying forecast probabilities for dissemination to forecast users.

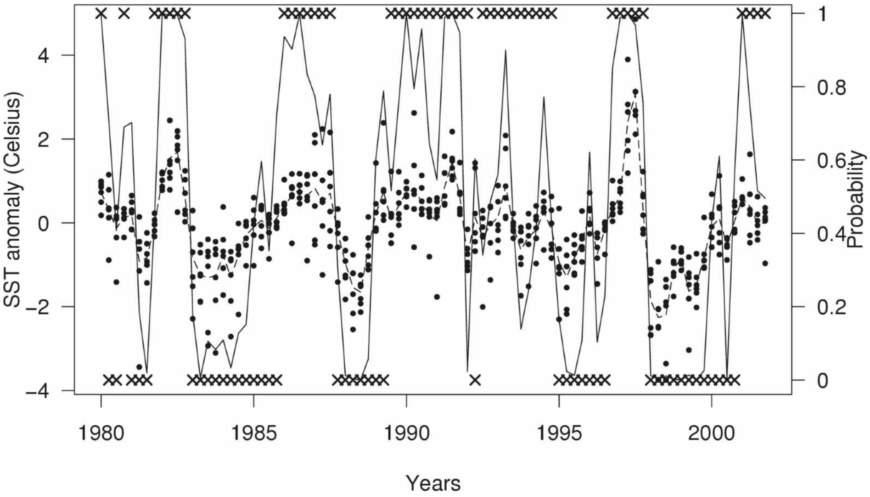

FIG. 1. Example of the ensemble forecasts and binary observations at one grid point located at 150°W in the central equatorial Pacific.

This could be considered to be a recalibration of the probabilities using a staircase function with a finite number of discrete steps. The Brier score REL − RES + UNC is what the user would obtain with the simplified probabilities compared to the full Brier score REL − GRES + UNC that would be obtained by the forecaster before simplification. The difference between GRES and RES therefore measures how much the simplification has degraded the Brier score of the forecast product.

3. Example: Multimodel ensemble SST forecasts

This section illustrates the Brier score decomposition using an example of multimodel ensemble forecasts of sea surface temperatures in the tropical equatorial Pacific produced by the coupled multimodel ensemble system as part of the European Union's Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) project (Palmer et al. 2004). The example is the same as that described in detail in Stephenson et al. (2005). The binary event was defined by whether a sea surface temperature (SST) anomaly was greater than zero. The probability forecasts were constructed parametrically by fitting a normal (Gaussian) distribution to the seven ensemble mean anomaly forecasts from the seven DEMETER coupled models, and then calculating the area under the normal density for values greater than zero. The forecasts can take any value between 0 and 1. The forecasts were issued a total of 88 times at 0000 UTC on the first day of February, May, August, and November from 1980 to 2001. A time–longitude section of forecasts at 56 gridpoint locations from 140°E to 82.5°W along the equator is shown in Fig. 3f of Stephenson et al. (2005).

Figure 1 shows an example of the observed binary event (the crosses; right-hand scale) and the probability forecasts (solid line; right-hand scale) for one grid point at 150°W in the central equatorial Pacific. The probability values were obtained by fitting normal distributions to the ensemble mean anomaly forecasts of the seven models (solid dots; left-hand scale). The mean of this distribution is given by the mean of the seven model forecasts (dashed line; left-hand scale). It can be noted that there is generally a good positive association between the probability forecasts and the observations.

The Brier score and its components were calculated by pooling events over all 56 gridpoint locations and the 88 different dates resulting in a total of 4928 binary observed events: 2472 zeros and 2456 ones. The overall Brier score computed without any binning was found to be 0.19, which is less than the expected Brier score, BS $= s_o^2 = \overline{o}(1 - \overline{o}) = 0.25$, one would obtain if one had issued a constant climatological probability of $f = 0.5$ for random events with a relative frequency of $\overline{o} = 0.5$. The multimodel forecasts are therefore more skillful for this sample of events than are climatology forecasts.

The components of the Brier score were computed by partitioning the probability forecasts into sets of equally spaced probability bins covering the unit interval with decreasing bin widths of 1.0, 0.5, 0.2, and 0.1 for increasing number of bins of 1, 2, 5, and 10, respectively. Figure 2 shows how these components vary with the number of bins. The sum of the three components (dashed line) exceeds the unbinned Brier score of 0.19

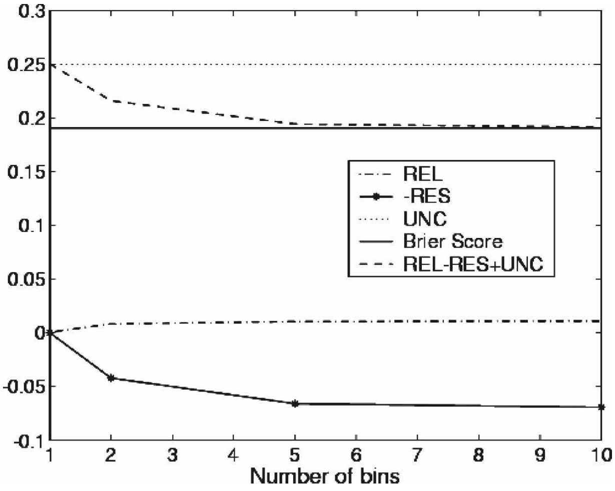

FIG. 2. Bin dependence of the unbinned Brier score (solid line) and its three traditional components: uncertainty (dotted line), reliability (dotted–dashed line), and negated resolution (line with asterisks).

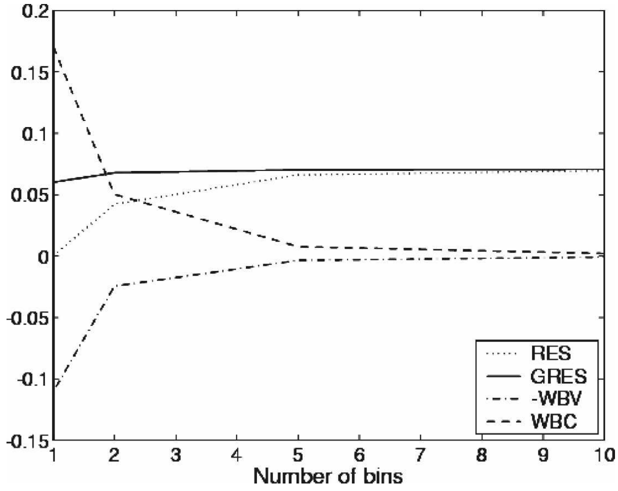

FIG. 3. The generalized resolution component GRES (solid line) defined as the sum of the traditional resolution component RES (dotted line) and the negated within-bin variance (dotted–dashed line) and the within-bin covariance (dashed line).

(the solid line) by up to 25% when calculated using a small number of bins. Unlike the uncertainty component (the dotted line) that depends solely on the frequency of the observed event, the negated resolution component (the line with asterisks) and the reliability component (the dotted–dashed line) of the Brier score both depend on the number of bins. The resolution component is zero when partitioning into only into one bin (by definition) but then increases to a larger value as the number of bins is increased. The reliability component is zero for one bin (no unconditional bias in this particular anomaly example) but then also increases to a larger value as the number of bins increases. The increase in the reliability component with more bins was also demonstrated in Fig. 2 of Atger (2004) and is to be expected from general mathematical considerations (Candille and Talagrand 2005). It should be noted, however, that the increases do not have to be strictly monotonic due to the presence of random sampling variations. The rate of increase in the components will depend on details such as the number of forecasts in each bin. The overestimation of the Brier score for small numbers of bins is due to the overestimation in the negated resolution component being greater than the overestimation in the reliability component. It is not mathematically clear yet whether this overestimation is a universal characteristic valid for all examples. However, for unconditionally unbiased forecasts both the reliability and the resolution components become zero when the number of bins equals unity, and hence the estimated Brier score reduces to the uncertainty term, which is greater than the unbinned Brier score providing the forecasts have some skill.

Figure 3 shows the negated WBV (the dotted–dashed line) and the WBC (the dashed line) components as a function of the number of bins. Both terms tend to zero with an increasing number of bins and would be exactly zero if one had stratified on the probability values that had been issued. In this example where there are many rather skillful forecasts, the within-bin terms are substantial compared to the reliability and resolution terms when five or fewer bins are used. However, for systems with less skill and fewer forecasts, the terms could remain substantial even when using more bins. The importance of the within-bin terms can be seen in Fig. 3, which compares the new generalized resolution component (the solid line) to the traditional resolution component (the dotted line). The generalized resolution component is much less sensitive to the choice of bin size than is the traditional definition of resolution.

4. Conclusions

This study has shown that the three-component Brier score decomposition is only valid if one stratifies on all issued values of forecast probability. If one first partitions the probabilities into bins before stratifying (as is often done to produce reliability diagrams), then it is necessary to consider a further two components to account for within-bin variation and covariation. The two within-bin components can be added to the resolution component to define a generalized resolution component that is less sensitive to the choice of bin width. However, because the Brier score and uncertainty terms are independent of bin width, as bin width decreases, generalized resolution must increase in order to compensate for the increase in reliability that has been demonstrated by Atger (2004) and Candille and Talagrand (2005). The difference between the generalized resolution and resolution components indicates how much the Brier score increases by binning the forecasts—it provides a measure of the loss of skill caused by binning the probability forecasts.

One can consider binning to be a crude form of nonparametric smoothing. Smoothing has the advantage of improving estimates of conditional averages and so can give estimates of Brier score components that have less sampling uncertainty than if one stratified on each single value of issued probability. However, smoothing can also introduce bias and so one needs to develop good statistical modeling approaches. Atger (2004) addressed the problem by parametric modeling of the reliability curve based on binormal linear fits to ROC curves. It would be of interest to develop more flexible nonparametric approaches for estimating the reliability curves (and hence the Brier score components), which also included point-wise confidence intervals.

Acknowledgments. Much of this work was done while ITJ was funded by an NERC Environmental Mathematics and Statistics Discipline-Bridging award (NER/T/S/2003/00126). CASC was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through Grant 200826/00-0 and by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), Processes 2005/05210-7 and 2006/02497-6. We also wish to thank the reviewers and our colleagues, especially Dr. Chris Ferro, for their useful comments that helped to improve this note.

REFERENCES

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.,* **131,** 1509–1523.

——, 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.,* **130,** 627–646.

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

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.,* **131,** 2131–2150.

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

Ferro, C. A. T., 2007: Comparing probabilistic forecasting systems with the Brier score. *Wea. Forecasting,* **22,** 1076–1088.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 254 pp.

Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **17,** 173–191.

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

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

——, 1986: A new decomposition of the Brier score: Formulation and interpretation. *Mon. Wea. Rev.,* **114,** 2671–2673.

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

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.,* **2,** 191–201.

Stephenson, D. B., C. A. S. Coelho, M. Balmaseda, and F. J. Doblas-Reyes, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus,* **57A,** 253–264.