The Correspondence Ratio in Forecast Evaluation

DAVID J. STENSRUD
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

MATTHEW S. WANDISHIN
The University of Arizona, Tucson, Arizona

(Manuscript received 17 May 1999, in final form 30 March 2000)

ABSTRACT

The correspondence ratio is developed to evaluate output from an ensemble of numerical weather prediction models. This measure is a simple extension of the threat score, or critical success index, to more than two fields and is used to measure the divergence of the forecast fields. The ratio is compared with two commonly used measures: the anomaly correlation, and the mean square error. Results indicate that the correspondence ratio is sensitive to the bias and, when calculated for several threshold values, can provide information beyond that supplied by the mean-square error and anomaly correlation measures. The correspondence ratio is particularly useful in evaluating discontinuous fields, such as precipitation. While no one measure can provide a complete assessment of forecast success, this ratio provides useful information that can increase our understanding of model forecast quality.

1. Introduction

Evaluating the quality, and/or error growth, of predictions from numerical weather forecasts is a difficult problem, since the observational and forecast distributions are multidimensional, yet a simple one-dimensional summary measure is often desired (Wilks 1995). While many of these scalar measures have been found to be useful (Panofsky and Brier 1958; Murphy and Duan 1985), some of the information contained within the data is lost when the dimensionality is reduced. Clearly no single, simple measure of forecast quality gives a complete measure of forecast success (Murphy and Winkler 1987; Murphy 1995). Yet it is evident that reducing the dimensionality of the data is necessary, since it is impossible to make sense of the full dataset.

This problem of data reduction is particularly acute when examining output from an ensemble of numerical forecasts. An ensemble is a set of model forecasts that are valid over the same forecast period, but use slightly different initial conditions and/or model configurations (Mullen and Baumhefner 1988, 1989; Tracton and Kalnay 1993; Molteni et al. 1996; Houtekamer and Lefevre 1997; Buizza 1997; Stensrud et al. 1999). One of the challenges of ensemble forecasting is developing methods to evaluate quantitatively the information content of the ensemble. Typically this has been done by calculating the mean-square error (mse), or one of its variants, of the ensemble mean (Hollingsworth et al. 1980; Kalnay et al. 1990; Stensrud et al. 1999), the variance of the ensemble members from the ensemble mean value (Arpe et al. 1985; Mullen and Baumhefner 1988, 1989), the anomaly correlation (Miyakoda et al. 1972; Hollingsworth et al. 1980; Arpe et al. 1985; Kalnay et al. 1990), rank histograms (Hamill and Colucci 1997, 1998), reliability diagrams (Wilks 1995; Hamill and Colucci 1998), various threat scores (Molteni et al. 1996; Du et al. 1997), and the relative operating characteristic curves (Mason 1982, 1989). Each of these measures provides a different perspective on the quality of the forecasts.

Two of the more commonly used measures in evaluating ensemble output are the anomaly correlation (AC) and the root-mean-square error (rmse). As discussed by Murphy and Epstein (1989), both of these measures have their strong and weak points. The AC is a good measure of the linear association between two fields. In contrast, the rmse ignores the linear association between two fields, but is a good measure of accuracy, the average degree of correspondence between two fields (Murphy 1995). Therefore, both measures are often used to evaluate forecast performance.

Another measure that has been used increasingly in evaluating the output from probabilistic ensemble forecasts is the relative operating characteristic (ROC) (Ma-
The ROC evaluates probabilistic forecasts through a calculation of the hit rate and false alarm rate for each value of decision probability. As discussed by Mason (1982), the ROC can be used to compare forecasts for dichotomous predictands in any form, whether probabilities or binary yes/no decisions. The ROC represents a general method for assessing forecast quality and comparing probabilistic and deterministic forecasts.

As an example of the difficulties encountered when a single measure is used to evaluate forecast quality, precipitation forecasts from 43 cases, utilizing the National Centers for Environmental Prediction (NCEP) short-range ensemble system output, are used to calculate the root-mean-square differences (rmsd) and AC values (see the appendix). Here the rmsd is calculated instead of the mse since we want to examine the dispersion generated by the ensemble and are not interested in the model bias. Hence, the observed precipitation amounts are not incorporated into these calculations. The values of rmsd and AC are computed as the mean value over all possible ensemble member pairs. This ensemble consists of 10 members of the 80-km version of the NCEP Eta Model (Janjić 1994) and 5 members from the 80-km regional spectral model (Juang and Kanamitsu 1994). Six of the Eta members have initial conditions created from slightly different in-house analyses, while the remaining four members have initial conditions created from two positive and two negative bred perturbations (Toth and Kalnay 1993). The regional spectral model members are initialized from one analysis member and the four bred perturbation initial conditions from the Eta Model ensemble. The reader is referred to Hamill and Colucci (1997, 1998) and Stensrud et al. (1999) for further information.

Results indicate that the values of rmsd from the 12-h ensemble precipitation forecasts grow very slowly with time (Fig. 1). In contrast, the values of AC decrease slowly with time, and remain above 0.68 throughout the 48-h period (Fig. 1). The rmsd curve can be interpreted in two very different ways. One interpretation is that the spread in the precipitation forecasts is growing linearly with time at a very slow rate. The other interpretation is that the precipitation forecasts have reached diffusion saturation during the first 12 h of the forecasts, since the growth in the differences between the forecasts is small during the remaining 36 h (see Lorenz 1965). The AC values support the first interpretation, with the values only declining from 0.8 to 0.68 over the 48-h period.

Unfortunately, neither the rmsd nor the AC is well designed for evaluating a discontinuous field like precipitation, yet precipitation is one of the most important of all model parameters and cannot be neglected. The difficulty with precipitation forecasts is that large areas have values of zero, punctuated by small regions with positive values. Large regions with no precipitation can dominate the calculations of rmsd and AC, tending to reduce their variability with time, yet rewarding the forecast for producing no precipitation in the correct places is desirable. Further interpretation of these results requires additional information, which can be supplied by the correspondence ratio.

The correspondence ratio, a simple measure of the agreement between two or more fields that represents an extension of the threat score, or critical success index (Donaldson et al. 1975; Schaefer 1990; Wilks 1995) to more dimensions, is developed in section 2 and used with a range of threshold values in section 3. The relationship between this measure and the AC and mse is explored in section 4, followed by a discussion in section 5. While this new measure suffers from the same problems in data reduction as the AC and mse, it has several characteristics that may be useful in attempting to assess model performance, especially when examining the predictability of model ensembles from discontinuous fields.

2. The correspondence ratio

In illustrating concepts in probability, the Venn diagram is often introduced. This geometric diagram illustrates the relationships among events in a sample space (Fig. 2) and further can be used to indicate the probability of the union of intersecting events (see Wilks 1995). The correspondence ratio is an attempt to reproduce numerically in a single number the information contained in the simple Venn diagram.

The correspondence ratio (CR) is defined as the area of the intersection (I) of all specified field values divided by the area of the union (U) of these same specified field values, such that

\[
CR = \frac{I}{U} = \frac{\sum_{i=1}^{n} f_{i,1} \cap f_{i,2} \cap \ldots \cap f_{i,m} \cap o_i}{\sum_{i=1}^{n} f_{i,1} \cup f_{i,2} \cup \ldots \cup f_{i,m} \cup o_i},
\]
where \( f_{m,i} \) represents the value of the \( m \)th forecast, or ensemble member, at the \( i \)th point; \( o_i \) represents the value of the observation at the \( i \)th point; and \( i = 1, n \) where \( n \) is the total number of data points. The CR is calculated for a chosen threshold value of the fields \( f \) and \( o \) (either a single value or a range of values). If one only examines the correspondence between ensemble members, then the intersection with \( o_i \) is neglected in (1). When calculating the correspondence between two fields, the CR is identical to the well-known threat score (Donaldson et al. 1975; Schaefer 1990; Wilks 1995) and can be viewed as a simple extension of the threat score to more dimensions. Like the threat score, the CR can be calculated for varying thresholds to determine the largest value of a given parameter for which the model forecasts have some skill. In addition, the evolution of the CR for one particular threshold value provides an estimate of predictability error growth and can be used to evaluate the divergence of the forecast fields with time.

As an example, imagine a 12-h rainfall forecast from two numerical models. Both show broad areas of precipitation occurring in association with a midlatitude cyclone (Fig. 3). These regions of precipitation are similar at the beginning of the forecast, but diverge with time as the differences in the two models and their respective initializations produce growth in different directions in phase space. If we define the specific forecast for the CR to be total rainfall of greater than or equal to 1 mm in 12 h, the CR is just the area of the intersection of these two model precipitation forecasts divided by the area of their union. Thus, the CR is equivalent to the threat score for these two forecasts. The CR is likely close to 1 at the beginning of the forecast and decreases as the forecast time increases and the spatial agreement between the two forecast fields is reduced. The added advantage of the CR is that this calculation can be produced easily for any number of ensemble members and is not limited to just two fields as illustrated. The threshold values of the selected parameters can be easily changed to allow the CR to measure the agreement between the two fields over the entire range of amplitude.

The CR is sensitive to bias in the forecasts. If the models overforecast the area being evaluated, yielding a bias value greater than 1, then the CR decreases since the area of the union of all the forecast/observed regions is larger than area of the intersection of the regions. Similarly, if the models underforecast the area, yielding a bias value less than 1, then the CR again decreases, since the area of the intersection of the forecast/observed regions is smaller than the area of the observations. Thus, the CR has a maximum value of 1 and is bounded below by 0. Results with simple analytic functions, discussed further below, indicate that the CR behaves similarly to (AC), but is not as adversely affected when using discontinuous fields.

Unlike calculations of the ROC (Mason 1982) and the probabilistic threat score (E. Kalnay 1999, personal communication), the CR as used here is not determined using the ensemble probabilities. However, a probabilistic CR can be defined (see the appendix). This probabilistic CR would allow for a more complete examination and exploration of the ensemble data. The evolution of the values of the probabilistic CR and the ROC then could be compared. However, we choose to illustrate the utility of the CR without using the full probabilistic information. Future studies may want to in-
investigate the whole range of the ensemble data with the probabilistic CR.

In addition to providing information on the divergence of the forecast fields, the CR also provides information on the largest shared amplitude of the fields. Results further suggest that the CR is particularly useful when the model fields are discontinuous. The CR is related to the ROC in that the hit rate used in the ROC calculation increases as the CR increases, but the two measurements examine different aspects of forecast quality.

a. CAPE and CIN verification

Thirty forecasts of convective available potential energy (CAPE) and convective inhibition (CIN) (Colby 1984), from the NCEP short-range ensemble forecast system, are used in calculating the CR for various threshold values. The Eta Model analyses valid at the 12–48-h forecast times are used for verification. The ensemble does best in predicting the lower values of CAPE, with CR values above 0.4, and has increasing difficulty in predicting the larger values of CAPE that can be important in forecasting severe weather events (Fig. 4). However, the largest CR value of 0.5 for the 500 J kg\(^{-1}\) CAPE threshold at the 12-h forecast time still shows a large uncertainty in the location of the predicted regions of instability. Also note that while the CR values remain nearly constant over the 48-h time period for most thresholds, the CR decreases by a factor of 2 for the 3000 J kg\(^{-1}\) threshold, indicating a doubling in the uncertainty. These results suggest that the distribution of CAPE is not predicted well by this ensemble, although an analysis of the full ensemble probabilities may yield a different outcome.

Unlike the CAPE field, the CIN field shows a clear diurnal cycle (Fig. 5). Since 26 out of the 30 forecasts are initialized at 1200 UTC, the increase in the CR values at the 24- and 48-h times indicates that the prediction of CIN is better in the morning than in the evening. This may be due to the increase of CIN during the night as the boundary layer cools. Again, the CR values decrease as the threshold value of CIN increases, indicating the greater difficulty of forecasting high CIN areas correctly. In addition, the CR values for CIN are smaller than those calculated for CAPE. This suggests that predicting CIN is more difficult than predicting CAPE, likely owing to the CIN being even more sensitive to the boundary and inversion layer evolutions.

b. Predictability of precipitation fields

The precipitation forecasts from the same 43 cases of NCEP short-range ensemble system output, as described in section 1, are used to calculate the CR values for the 12-h precipitation totals. Only the model forecasts are used in these calculations, since we are examining the model dispersion characteristics. Plots of the areas of intersection and union from one case illustrate the decrease in the CR that occurs as the area of intersection decreases in size and the area of union increases in size (Fig. 6). Results using precipitation threshold values of 1, 2, and 5 mm from all the cases indicate that the forecast fields diverge substantially with time (Fig. 7), with the larger threshold values creating faster divergence in the fields. When interpreted in conjunction with the rmsd calculations (see Fig. 1), these results suggest that the precipitation forecasts have reached spread saturation, since it is difficult to accept that the divergence of the fields doubles (i.e., the CR values are halved) while the precipitation amount forecasts are still in agreement.\(^1\) This implies that this particular model is unable to accurately predict 12-h precipitation totals, yet is able to predict the precipitation region out to at least 36 h before the forecast field divergence doubles. This result is consistent with our qualitative expectations of model forecast error growth, which suggests that models can forecast the general regions of precipitation

\(^1\)This conclusion has been confirmed using hourly precipitation output from a different model ensemble.
out to 36 h but have little skill in forecasting precipitation amounts. The values of CR provide information beyond that provided by the mse and AC, which only indicated a slow increase in spread with time and provide no information on the time over which the field divergence doubles.

As the threshold value used in the calculation is increased, the value of the CR decreases and the doubling time, or predictability time limit, also decreases. This is consistent with our understanding of model accuracy, since it is more difficult to predict the larger precipitation amounts. The threshold values could be increased further to measure the amount of agreement for higher precipitation amounts. Since the CR requires the definition of the threshold value for the chosen parameter to be used in the calculation, it can be calculated over a range of values as is done with the threat score. This flexibility allows for the evaluation of specific forecast features, instead of a global measure of forecast skill as found with the AC and rmse. A range-based application of the CR is now illustrated that can selectively examine specific model forecast features.

3. 500-hPa height patterns

The spaghetti diagram (Tracton and Kalnay 1993) is often used to display medium-range ensemble output for forecasting purposes. At the 500-hPa level, this diagram is a composite of a given height-level contour from all of the individual ensemble members. It is an effective tool for displaying ensemble data, since it succinctly illustrates the spread of the ensemble members and how the spread changes over the time period of the forecast. However, it is not a quantitative measure of the correspondence between the forecasts. In contrast, the CR can be designed to quantify this information by choosing a height range, instead of a threshold value, for defining the areas of union and intersection.

To illustrate the ability of the CR to quantify some of the information content in the spaghetti diagram, the 5600 ± 30 m height interval on the 500-hPa surface is chosen to define the areas of union and intersection, and the CR is calculated as shown in section 2. Results from the 15-member NCEP short-range ensemble forecasts initialized at 1200 UTC 2 December 1997 show that even at the initial time substantial differences are present in the height fields (Fig. 8a), leading to a CR value of 0.42 (Fig. 9). These differences grow with time (Figs. 8, 9), producing a CR value near 0 at 48 h. While the plots of union and intersection do not show the individual contour lines, an envelope of these contours is produced, yielding information similar to that provided by the spaghetti diagrams. In addition, the correspondence between the individual ensemble members is quantified using the CR, and could be used to compare different forecasts or assist in defining the forecast confidence for specific parameters of interest through an examination of the divergence of the forecast fields over time.
4. Relationships with AC and mse

Results suggest that, in many ways, the CR behaves similarly to the AC, a commonly used measure of association that operates on pairs of gridpoint values from two-dimensional fields (Wilks 1995). To explore the relationship between the AC, mse, and CR (see the appendix for definitions of AC and mse), we define two two-dimensional anomaly fields \( f_1, f_2 \) as trigonometric functions, where one field is held fixed and the other has added phase and amplitude differences, such that

\[
f_1 = \cos(x) \sin(y) \quad f_2 = B \cos(x + \varphi) \sin(y),
\]

where \( B \) is an amplitude, \( \varphi \) is the phase difference in the \( x \) direction, \(-\pi \leq x \leq \pi\), and \( 0 \leq y \leq \pi\). The curves of AC, CR, and mse are examined for different values of \( 0 \leq \varphi \leq \pi \) and \( 0.25 \leq B \leq 1 \). Results show that the CR and AC values follow similar trends as the phase difference of the two fields is increased (Fig. 10). However, the AC becomes negative after the phase difference increases past \( \pi/2 \), while the value of CR remains positive for the smaller threshold values until the phase difference approaches \( \pi \) and the value of CR approaches zero. Also note that as the threshold value for the CR calculation is increased, the CR curve goes to zero more rapidly, since there are fewer points where both of the anomaly fields are greater than the specified threshold value. For \( B = 1.0 \), the CR using a threshold value of 0.75 is nonzero for phase differences of less than \( \pi/2 \) (Fig. 10b), indicating that the two anomaly fields both have amplitudes of at least 0.75.

As the amplitude of the second anomaly field is lowered, the values of CR change (Figs. 10c–e). This change in amplitude has no effect on the values of AC, which remain identical for all values of \( B \) and only change as the phase is varied, as also shown by Brier and Allen (1951) and Murphy and Epstein (1989). However, in more realistic cases, for which the anomaly fields are often complex, the AC is sensitive to amplitude differences (Arpe et al. 1985). As the value of \( B \) decreases, some of the CR curves become zero over the entire range of phase differences. Indeed, a progression occurs in which the CR curves go to zero first for the largest threshold value (0.75), and then for the successively smaller threshold values (0.5, 0.25) as the difference in the amplitudes of the two anomaly fields is increased (i.e., \( B \) is decreased). In addition, the CR curve for the threshold value of 0.05 moves slightly lower as \( B \) is decreased, and does the maximum value of CR. Therefore, the values of CR are influenced by differences in amplitude between the two anomaly fields, and the largest threshold value for which CR is nonzero provides an indication of the largest shared amplitude of the anomaly fields for this example. This relationship applies for phase differences less than \( \pi/2 \). Beyond this...
Fig. 10. Idealized calculations of (a) anomaly correlation, and (b)–(e) correspondence ratio vs phase differences (\(\phi\)) from the analytic two-dimensional fields. The effects of amplitude differences are also examined in (b)–(e), where (b) \(B = 1.00\), (c) \(B = 0.75\), (d) \(B = 0.50\), and (e) \(B = 0.25\). In (b)–(e), the thick black line indicates a threshold value of 0.05, the thick dashed line indicates a threshold value of 0.25, the thin dashed line indicates a threshold value of 0.50, and the dash–dot line indicates a threshold value of 0.75.
Fig. 11. Idealized calculations of mean-square error vs phase difference ($\phi$). The effects of amplitude differences are also examined, where the thick solid line is for $B = 1.00$, the thick dashed line is for $B = 0.75$, the dashed line is for $B = 0.50$, and the dash-dot line is for $B = 0.25$.

point, the differences in the fields could be produced by either phase or amplitude differences.

The curves of mse are substantially different from those of the AC or CR in that they increase as the phase difference between the two anomaly fields is increased (Fig. 11). This is an indication that the average agreement between the two fields is decreasing as the phase difference increases, such that the overall trend of the mse is opposite to the trends in the AC and CR. Within the range $\pi/5 \leq \phi \leq 2\pi/5$ the curves of mse from the different values of $B$ cross over one another. This indicates that it is not possible to distinguish with certainty whether or not there is any difference in amplitude between these two fields by using the mse for AC values less than approximately 0.85. Other measures, such as elements of the decomposition of the mse (Takacs 1985; Murphy and Epstein 1989) or using wavelets (Briggs and Levine 1997), are needed to make this determination. Similar information can also be extracted from the CR when using various threshold values, suggesting that the CR may be a useful parameter for the evaluation of model output.

5. Discussion

We have shown that the CR is a simple extension of the threat score to more dimensions and can be used to measure the divergence of ensemble forecast fields. This ratio requires the definition of either a threshold value, or a range of values, for computation. Choosing between a threshold value, or a range of values, depends upon the field that is being examined. For precipitation forecasts, or anomaly fields, a threshold value is recommended, whereas for 500-hPa height fields, a value range could be used. This flexibility is an advantage, since one does not have to know the underlying climatological mean, as is needed to calculate the AC. Results indicate that the CR is particularly useful when examining the divergence of discontinuous fields, such as precipitation, for which the evaluation of the mse and AC calculations are largely controlled by the large number of forecasts of no precipitation. A probabilistic version of the CR, defined in the appendix, would allow an even greater exploration of the ensemble dataset.

Varying the threshold values also allows one to estimate the amplitude differences of the fields being evaluated, since as the amplitude differences increase the CR values decrease for larger threshold values. The threshold value for which the CR first goes to zero is an estimate of the largest shared amplitude of the fields, assuming the phase difference of the two fields is less than $\pi/2$. It also can provide information beyond that supplied by the mse, since it can be designed to explore particular aspects of the model forecasts. The CR is related to the ROC in that the hit rate used in the ROC calculation increases as the CR increases, but the two measurements examine different aspects of forecast quality. The CR also is more versatile since it can be used to compare different forecasts apart from their verification, as is often done in predictability studies.

While the CR cannot be clearly linked to the mse, as occurs with the AC (Murphy 1988), it offers a different perspective on the quality of the forecasts and appears to help in interpreting the values provided by other measures of forecast quality. While no one measure can provide a complete assessment of forecast success, this ratio provides useful information that can increase our understanding of model forecast quality.

Acknowledgments. The authors thank Drs. Barbara Brown and Harold Brooks for their comments on an earlier version of this manuscript, and Dr. Steven Mullen for his encouragement and many enlightening discussions. Two anonymous reviewers also helped to greatly improve this presentation and their efforts are appreciated. This work was supported in part by the NSF under Grants ATM-9424397, ATM-9501532, and ATM-9612487.

APPENDIX

Statistical Measures

The mean square error is defined as

$$\text{mse} = \frac{1}{M} \sum_{m=1}^{M} (x_m - y_m)^2,$$

(A1)

where $x$ and $y$ represent two different fields and $M$ is the total number of data points. The rmse is the square
root of the mse. Similarly, the rmsd is also the square root of the the mse when both \( x \) and \( y \) are model fields, thereby representing the difference between model fields and not the difference between model and observations. When calculating the rmsd from an ensemble, an average of all the rmsd values is determined from each possible pair of ensemble members.

The AC is defined as

\[
AC = \frac{\sum_{m=1}^{M} [(y_m - C_m)(x_m - C_m)]}{\sum_{m=1}^{M} (y_m - C_m)^2 \sum_{m=1}^{M} (x_m - C_m)^2}^{1/2}, \quad (A2)
\]

where \( x \) and \( y \) represent different two-dimensional representations of a variable, \( C \) represents the climatological average of the given variable for each grid point, and \( M \) is the total number of grid points. Often \( x \) is defined as the observed field at the model \( y \) forecast time, although \( x \) could just as easily be a different model forecast valid at the same time. Note that \(-1 \leq AC \leq 1\), and that a value of \( AC = 0.5 \) has been found to correspond to a climatological skill score of zero (Murphy and Epstein 1989). Murphy and Epstein (1989) further indicate that the square of the AC should be interpreted as a measure of potential rather than actual skill.

When calculating the AC from an ensemble, an average of all the AC values is determined from each possible pair of ensemble members.

Murphy (1988) shows that the mse can be decomposed into terms that include the AC, such that

\[
\text{mse} = (\bar{x} - \bar{y})^2 + \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y r_{xy}, \quad (A3)
\]

where the overbar represents a mean value, \( \sigma \) is the sample standard deviation, and \( r_{xy} \) is the sample covariance between \( x \) and \( y \). If \( x \) and \( y \) are anomaly fields with zero sample means, then \( r_{xy} \) is the anomaly correlation coefficient. Therefore, as the AC is increased the mse should decrease, with the magnitude of the decrease depending upon the sample standard deviations. Note that the mse \( \geq 0 \), with equality indicating that the two fields are identical.

A probabilistic version of the CR (CR\(_p\)) is defined as the area associated with ensemble probabilities exceeding a specified threshold value for a specific forecast aspect divided by the area of the union of the same specific forecast aspect. Thus, the denominator in the probabilistic CR is the same as in (1), while the numerator is based upon the ensemble probabilities. This is quantified as

\[
\text{CR}_{p} = \frac{\sum_{i=1}^{n} \left\{ \left( \frac{1}{m} \sum_{j=1}^{m} d(f_{j,i}) \right) \geq P \right\} \cap o_i}{\sum_{i=1}^{n} f_{1,i} \cup f_{2,i} \cup f_{3,i} \cup \cdots \cup f_{m,i} \cup o_i}, \quad (A4)
\]

where

\[
d(f_{j,i}) = \begin{cases} 
1 & \text{for } f_{j,i} \geq \text{threshold} \\
0 & \text{for } f_{j,i} < \text{threshold}
\end{cases} \quad (A5)
\]

and \( f_{j,i} \) represents the value of the \( j \)th forecast, or ensemble member, at the \( i \)th point; \( o \) represents the value of the observation at the \( i \)th point; \( i \) is the total number of data points; \( m \) is the number of ensemble forecasts; and \( P \) is the threshold ensemble probability value chosen. The probabilistic CR is again calculated for a chosen threshold value of the fields \( f \) and \( o \) (either a single value or a range of values). The CR\(_p\) is shown graphically in Fig. 8. The denominator is indicated by the area of light shading, the numerator of the CR\(_p\) for the 66.6% level is defined by the area of medium shading (the region where 10 out of 15 members intersect), and the numerator of the CR\(_p\) for the 100% level is defined by the area of dark shading. The original CR is identical to the CR\(_p\) for the 100% probability level. The CR\(_p\) values range from 0 to 1.

Calculations of the probabilistic CR\(_p\) from a 15-member ensemble of Eta and Regional Spectral Model (RSM) model forecasts show that the time over which the CR\(_p\) is halved increases as the probabilities used to define the numerator decrease (Fig. A1). Similar calculations, when combined with observational data, could be used to explore the behavior of an ensemble system for a wide variety of probability values.

**REFERENCES**
