Measuring the Ensemble Spread-Error Relationship with a Probabilistic Approach: Stochastic Ensemble Results

Eric P. Grimit* and Clifford F. Mass
Department of Atmospheric Sciences
University of Washington, USA

Submitted to Monthly Weather Review
June 29, 2005

Revised March 10, 2006

*Corresponding author: Department of Atmospheric Sciences, University of Washington, Box 351640, Seattle, Washington 98195-1640, USA.
Abstract

One widely accepted measure of the utility of ensemble prediction systems is the relationship between ensemble spread and deterministic forecast accuracy. Unfortunately, this relationship is often characterized by spread-error linear correlations, which oversimplifies the true spread-error relationship and ignores the possibility that some end users have categorical sensitivities to forecast error. In the present paper, a simulation study is undertaken to estimate the idealized spread and error statistics for stochastic ensemble prediction systems of finite size. Under a variety of spread and error metrics, the stochastic ensemble spread-error joint distributions are characterized by increasing scatter as the ensemble spread grows larger.

A new method is introduced that recognizes the inherent nonlinearity of spread-error joint distributions and capitalizes on the fact that the probability of large forecast errors increases with ensemble spread. The ensemble spread-error relationship is measured by the skill of probability forecasts that are constructed from a history of ensemble-mean forecast errors using only cases with similar ensemble spread. Thus, the value of ensemble spread information is quantified by the ultimate benefit that is realized by end users of the probability forecasts based on these conditional-error-climatologies.

It is found that the skill of conditional-error-climatology forecasts based on stochastic ensemble spread is nearly equal to the skill of probability forecasts constructed directly from the raw ensemble statistics. The skill is largest for cases with anomalous spread and smallest for cases with near-normal spread. These results reinforce the findings of earlier studies and affirm that the temporal variability of ensemble spread controls its potential value as a predictor. Additionally, it is found that the skill of spread-based error probability forecasts is maximized when the chosen spread metric is consistent with the end user’s cost function.
1 Introduction

Ensemble numerical weather prediction (NWP) provides an approach for incorporating both initial condition (IC) and model uncertainty into the forecast process while automatically accounting for flow dependence (Ehrendorfer 1997; Palmer 2000). Given an ideal ensemble prediction system, one that accurately accounts for all sources of forecast uncertainty, the verifying truth should be indistinguishable from the members of the forecast ensemble (Anderson 1996; Hamill 2001). The spread of such an ideal forecast ensemble ought to provide an \textit{a priori} estimate of the forecast uncertainty: cases with large (small) ensemble spread should be associated with large (small) forecast uncertainty. The probability of specific weather events could be reliably specified from such perfect uncertainty forecasts, thereby allowing operational forecasters and other end users of weather information to determine their associated risk or reward and take appropriate action when needed (Katz and Murphy 1997).

Real ensemble NWP systems are naturally imperfect and they require statistical post-processing in order to generate calibrated probability forecasts for users. Nevertheless, uncalibrated probability forecasts that can be generated directly from the raw ensemble relative frequencies are skillful in many instances [e.g., see Figs. 1–4 of Arribas et al. (2005) and Fig. 6 of Buizza et al. (2005)]. Therefore, real ensemble prediction systems show the ability to capture some fraction of the true forecast uncertainty and this partial estimate can contain useful information.

As a simple measure of the utility of real ensemble spread, several researchers have evaluated whether or not ensemble spread gives a reliable prediction of the expected error for ensemble-mean or high-resolution control forecasts. The traditional approach to quantifying this mapping is to find the linear correlation between the ensemble spread and the deterministic forecast error over a large collection of separate cases (e.g., Kalnay and Dalcher 1987; Murphy 1988; Barker 1991; Buizza 1997). Pearson’s correlation coefficient (Wilks 2005) is typically calculated between time series of ensemble spread and deterministic forecast error, sometimes preceded by spatial averaging. The strength of the correlation indicates how well the ensemble spread and error data fit a linear model. Several definitions of ensemble spread and error have
been used, including root-mean-square (rms) metrics and anomaly correlations.\(^1\)

Regardless of the metrics used, the observed ensemble spread-error correlations described in the literature have been rather disappointing. For a range of forecast applications, from mid-tropospheric geopotential heights (e.g., Buizza 1997) to tropical cyclone tracks (e.g., Goerss 2000) and mesoscale convective precipitation (e.g., Stensrud et al. 1999), the ensemble spread-error distribution is usually highly scattered with linear correlation coefficients less than 0.6. In some cases, no relationship between ensemble spread and error is apparent (e.g., Hamill and Colucci 1998). Larger (up to 0.7) ensemble spread-error correlations have been reported with the aid of forecast bias correction (Stensrud and Yussouf 2003). In addition, several studies note that the strongest ensemble spread-error relationship is restricted to the subset of cases with anomalously large and small ensemble spread and that cases with near-normal ensemble spread show virtually no spread-error connection (Houtekamer 1993; Whitaker and Loughe 1998; Grimit and Mass 2002).

The reason for the generally low observed ensemble spread-error correlations is well explained with simple statistics. Houtekamer (1993) uses a two-parameter stochastic model to approximate the expected ensemble spread and ensemble-mean error statistics for an idealized ensemble prediction system that implicitly has infinite size. If the ensemble spread (, standard deviation) is assumed to vary log-normally and the error (, absolute error of the ensemble mean) is assumed to be normally distributed, then the ensemble spread-error correlation (\(\rho_{\sigma,|\bar{e}|}\)) can be written as an analytic function

\[
\rho_{\sigma,|\bar{e}|} = \sqrt{\frac{2}{\pi}} \cdot \frac{1 - \exp(-\beta^2)}{1 - \frac{2}{\pi} \cdot \exp(-\beta^2)},
\]

where \(\beta = \text{std}(\ln\sigma)\) is the standard deviation over time of the natural logarithm of the ensemble spread. In the upper limit of ensemble-spread temporal variability (infinite \(\beta\), the ensemble spread-error correlation asymptotes to \(\sqrt{\frac{2}{\pi}}\), a value just less than 0.8 [see Fig. 1 of Whitaker and Loughe (1998)]. Therefore, there exists an analytical limit to the strength of the spread-error correlation even for an ideal ensemble prediction system. Realistic values of \(\beta\) generally lie between 0.1 – 0.5 and depend on the season, forecast parameter, and lead time.

\(^1\)The chosen forecast error metric ought to be non-negative to obtain a spread-error correlation.
Using (1), these values of $\beta$ suggest an upper bound of $0.13 - 0.53$ for the ensemble spread-error correlation in most cases. Analysis of NCEP ensemble 250-hPa streamfunction forecasts over two winter seasons confirms that ensemble spread is most useful as a predictor of forecast error when it is anomalous and that ensemble spread-error correlations vary with the magnitude of the ensemble-spread temporal variability (Whitaker and Loughe 1998).

A factor not considered in ensemble spread-error correlation studies is the state-dependence of the metrics. Toth (1992) and Ziehmann (2001) argue that forecast error magnitude tends to be larger (smaller) when the forecast state is near climatological extreme (average) conditions. If this is true, then a value of ensemble spread that suggests above-normal forecast error in climatological average conditions may also indicate a below-normal forecast error when the forecast state is anomalous. Using spread-error linear regression alone can thus lead to ambiguous prediction of the relative size of forecast errors, unless the forecast state is also considered.

A categorical approach can take account of the forecast state, making the resulting forecast error predictions more robust to departures from climatological averages (Toth et al. 2001; Ziehmann 2001). In this approach, the forecasts are divided among equally-likely bins according to the climatological distribution of the relevant parameter. Ensemble spread can be measured by the degree to which the ensemble member population is distributed among these climatological bins. To achieve the same amount of ensemble spread during extreme situations as during normal conditions, larger deviations among the ensemble members are required. This is because the climatological categories are naturally wider near the tails of the distribution than near its center. In order for categorical ensemble spread to be considered large, the forecast ensemble members must be distributed across multiple climatological bins. Two primary measures of categorical ensemble spread are mode population and statistical (classical) entropy. Mode population ($M_{\text{mode}}$) is defined as

$$M_{\text{mode}} = \max(M_i); \quad i = 1, 2, \ldots, nbin,$$

where $M_i$ is the number of forecast ensemble members contained in bin $i$ and $nbin$ is the total
number of bins. Statistical entropy (ENT) is calculated as

\[
ENT = - \sum_{i=1}^{nbin} \frac{M_i}{M} \log \frac{M_i}{M},
\]

where \(M\) is the total ensemble size.

Toth et al. (2001) and Ziehmann (2001) show that categorical ensemble spread measures discriminate between forecast successes and failures\(^2\) better than continuous ensemble spread measures (e.g., standard deviation and variance). Built upon these results, the relative measure of predictability is an operational tool used at NCEP to forecast the probability of success and is based on the forecast ensemble member population of the climatological bin containing the ensemble mean. However, the categorical verification strategy used in these studies implicitly assumes that the users of interest do not have continuous cost functions\(^3\). This assumption is probably valid for most casual end users of weather forecast information, who might notice when the weather departs significantly from normal and who can tolerate small forecast errors. Certainly, there are other end users with cost structures that are continuous functions of forecast error (e.g., utility companies requiring power demand projections). It is not clear whether categorical ensemble spread metrics perform better than continuous spread metrics for end users with continuous cost functions. Idealized calculations need to be made, akin to the work of Houtekamer (1993), to verify the merits of a categorical strategy for different types of end users.

In search of a meaningful ensemble spread-error relationship, some researchers have examined the statistics of ensemble-mean forecast errors grouped by categories of increasing ensemble spread. Fraedrich et al. (2003) and Hamill (2003) note small increases in the sample means of absolute or rms error as one moves from small to large ensemble spread categories. Saetra and Bidlot (2004) find a strong linear correlation between the 90\(^{th}\) percentile of the absolute ensemble-mean errors and the average ensemble spread in each category. Atger (1999) and Wang and Bishop (2003) estimate the variance of ensemble-mean forecast errors in each ensemble variance category and plot the two variances together. Understanding this relationship

---

\(^2\)A successful (unsuccessful) forecast is defined as the event that the verifying bin is (is not) the same as the modal bin.

\(^3\)That is, the penalty for bad forecasts is not a continuous function of the forecast error.
between ensemble variance and ensemble-mean error variance is a fundamental component of
the present paper and forms the basis of the probabilistic spread-error analysis explored here.

In this paper, a non-dynamical, stochastic model is developed to establish the expected
ensemble-mean forecast error statistics for an ideal ensemble prediction system subject to
varying ensemble sizes, ensemble-spread temporal variances, and spread and error metrics.
A variety of ensemble spread and error metrics are tested to reflect the wide-ranging needs
of end users and to establish the sensitivity of the ensemble spread-error relationship to the
metrics utilized. It is argued that traditional ensemble spread-error correlations oversimplify
these relationships and ignore the inherent nonlinearity. The main objective of this work is
to encourage the use of ensemble spread-error analyses that consider the fully probabilistic
nature of the problem and to discourage the use of simple spread-error linear correlations.
A new method for measuring the ensemble spread-error relationship is introduced, in which
the forecast ensemble spread is evaluated according to its ability to predict the $distribution$ of
ensemble-mean forecast errors in a probabilistic sense. The benefit of using ensemble spread
to predict forecast error distributions is thus reflected by the positive impact it has on error
probability forecast skill. The potential impact of using ensemble spread to improve error
probability forecasts is established using stochastic model results and serves as a reference
for the evaluation of real ensemble prediction systems. The actual probabilistic analysis of
spread-error for a real ensemble prediction system is too rich a topic to be covered in full here,
and thus will be covered in a subsequent paper.

Section 2 describes the simple stochastic model of ideal ensemble spread and error data used
in this paper and a sensitivity analysis of these stochastic ensemble spread-error correlations.
In Section 3, the rationale for a fully probabilistic approach to this problem is developed and
a new probabilistic method of forecast error prediction with ensemble spread information is
explored. Section 4 concludes.
2 Spread-Error Correlation within a Stochastic Model Framework

2.1 A Simple Stochastic Model

Dealing with a simple model allows one to eliminate the complexities of high-dimensionality and computational expense and to focus on fundamental issues. Highly idealized simple model results provide estimates of the upper limit in performance of the real dynamical system that the simple model approximates. The idealized model approach followed here uses a modified version of the two-parameter stochastic model of Houtekamer (1993). The aim is to establish the expected univariate statistics of ensemble spread and error in the presence of an ideal ensemble prediction system and to examine the corresponding sensitivity to ensemble size, ensemble-spread temporal variability, and end-user cost function.

A more flexible version of the Houtekamer (1993) stochastic model is required to examine the sensitivity of traditional spread-error correlations and to test new probabilistic error forecasting methods. In particular, the Houtekamer (1993) stochastic model assumes an ideal ensemble of infinite size, where explicit knowledge of the forecast ensemble and verification values is not required. To accommodate finite ensemble sizes, explicit simulation of forecast ensemble and verification values is necessary. Knowledge of the explicit values also facilitates the examination of different spread and error metrics.

For each hypothetical forecast case \( n = 1, \ldots, N \), we do the following:

1. Generate a random value that represents the actual forecast uncertainty \( \sigma \) from a log-normal distribution with temporal variance \( \beta^2 \) [as in the original Houtekamer (1993) model]. This can be expressed mathematically as

\[
\ln(\sigma) \sim \mathcal{N}(\ln(\overline{\sigma}_F), \beta^2),
\]

where \( \mathcal{N} \) denotes a normal (Gaussian) probability density function (PDF). The climatological median forecast uncertainty is specified by \( \overline{\sigma}_F \).

2. Explicitly simulate an ideal forecast ensemble of size \( M \) by drawing \( M \) values from a Gaussian distribution with variance \( \sigma^2 \). For each forecast ensemble member \( x_m \), this can
be written as

\[ x_m \sim \mathcal{N}(Z, \sigma^2); \quad m = 1, 2, \ldots, M, \]

where \( Z \) is the mean of the ideal forecast PDF. Note that, \( Z \) itself is drawn from a Gaussian PDF with climatological mean \( Z_C \) and variance \( \sigma_C^2 \).

3. Draw the verification (\( y \)) from the same Gaussian distribution as the corresponding forecast ensemble to ensure statistical consistency, such that

\[ y \sim \mathcal{N}(Z, \sigma^2). \]

4. Calculate ensemble spread and forecast error metrics as appropriate for a given end user.

For example, the ensemble standard deviation (SD) is given by

\[ SD = \sqrt{\frac{1}{M - 1} \sum_{m=1}^{M} (x_m - \overline{x})^2}, \quad (4) \]

where

\[ \overline{x} = \frac{1}{M} \sum_{m=1}^{M} x_m, \]

and the absolute error of the ensemble mean (AEM) is written

\[ AEM = |\overline{x} - y|. \quad (5) \]

Pearson’s correlation coefficient is then calculated for the \( N \) values of spread and error. To ensure high statistical significance of the correlations in this simulation experiment, \( N \) is chosen to be large (\( N = 10^4 \)).

The ensemble size (\( M \)) and ensemble-spread temporal variability (\( \beta \)) are adjustable. The ensemble spread-error correlation can be shown to be independent of the remaining parameters of the model (\( \overline{\sigma}_F, Z_C, \) and \( \sigma_C \)) when continuous measures of spread and error are used\(^4\). However, categorical metrics of spread and error are very sensitive to the choices for these three parameters, since the actual values assumed by the forecast ensemble and verification must be organized into predefined bins. In our implementation, we choose \( \overline{\sigma}_F = 15 \) m, \( Z_C = 5460 \) m,

\[^4\text{In particular, Houtekamer (1993) argues that spread-error correlation in the } M \to \infty \text{ limit is independent of the choice for the climatological median spread } (\overline{\sigma}_F) \text{ and this seems to extend to the case of finite } M.\]
and $\sigma_C = 150$ m to emulate short-range forecasts of 500 hPa geopotential height ($Z$). These fixed parameters enable straightforward identification of the climatological bins for $Z$, which are defined simply from the quantiles of the Gaussian PDF with mean $Z_C$ and variance $\sigma^2_C$. Dramatically changing these settings when continuous spread and error metrics are chosen does not significantly alter the spread-error correlations, since the associated scatter diagrams are only linearly distorted.

The stochastic ensemble predictions are ideal because the ensemble values are drawn from a distribution that represents the actual forecast uncertainty, which is normally unknown. A built-in limitation of the stochastic ensemble comes from the finite sampling of the ideal forecast distribution, such that the potential for poor spread estimation is larger for smaller ensemble sizes. This mimics the situation found in real-world ensemble prediction, where the only available estimate of forecast uncertainty is given by the variation within the forecast ensemble at hand. It should also be noted that the number of degrees of freedom in this stochastic model is much smaller than the number of degrees of freedom in real NWP models. This may cause an overestimation of the potential spread-error correlation attainable with real ensemble prediction systems.

### 2.2 Spread-Error Correlation Sensitivity Analysis

An analysis of ensemble spread-error correlations is undertaken to examine their sensitivity to ensemble size, ensemble-spread temporal variability, and end-user cost function. Ideal ensemble forecasts are created using the simple stochastic model described in the previous section, with ensemble size varied from 2–50. Ensemble spread-error correlation experiments are repeated for ensemble-spread temporal variabilities ($\beta$) ranging from 0.1–0.9, in 0.2 increments.

Stochastic ensemble spread-error correlations using the sample standard deviation (SD, Eq. 4) and the absolute error of the ensemble mean (AEM, Eq. 5) are shown in Fig. 1. The SD-AEM correlation varies greatly and is a function of both ensemble size ($M$) and ensemble-spread temporal variability ($\beta$). SD-AEM correlations are significantly less than unity for all combinations of $M$ and $\beta$. The asymptotic limits of the SD-AEM correlations for finite-size ensembles are directly comparable to the theoretical limits for an infinite-size ensemble.
(from Eq. 1). For large $\beta (\beta = 0.9)$, the SD-AEM correlation asymptotes to about 0.7, as $M$ approaches 50. For moderate levels of ensemble-spread temporal variability ($\beta = 0.5$), SD-AEM correlations asymptote to about 0.5. SD-AEM correlations monotonically increase with $M$ and $M \geq 10$ appears to be sufficient within this idealized numerical setting. Asymptotic limits are attained faster for ensembles with higher ensemble-spread temporal variability.

To investigate the metric-dependent sensitivity of ensemble spread-error correlations, different ensemble spread and error metric combinations are tested (see Tables 1 and 2 for definitions). A variety of both continuous and categorical measures of ensemble spread and error are considered to reflect the wide-ranging criteria of end users. For brevity, only a few of the total set of metrics are presented here. A more detailed analysis can be found in Grimit (2004).

For continuous metrics of ensemble spread, attention is restricted to the sample standard deviation (SD). Although many other continuous metrics of spread could be applied, the simplicity and common use of SD is attractive and related measures are unlikely to provide substantially different results. Continuous error metrics in addition to AEM are explored, since AEM is only indicative of the deterministic performance of the forecast ensemble mean. An increasingly popular metric of error that can be applied to both discrete forecast ensembles and the continuous PDFs based on them is known as the continuous ranked probability score (CRPS; Wilks 2005, p. 302). In this simulation study, CRPS is used only to score Gaussian probability density forecasts. An analytic formula for the CRPS in this Gaussian case is given by Gneiting et al. (2005) and included in Table 2. In addition, two non-standard error metrics denoted by AAE and SDE are investigated because they incorporate ensemble-averages of the individual forecast ensemble member errors and seem likely to correlate better with ensemble spread. Note that, in practice, the opposite procedure would be undertaken, since one already has an established error metric on-hand and would need to search for a new spread metric to pair with it.

Categorical spread and error metrics are calculated by partitioning forecasts and verifications into predetermined bins. As in Toth et al. (2001) and Ziehmann (2001), mode population ($M_{\text{mode}}$, Eq. 2) and statistical entropy (ENT, Eq. 3) are defined using ten equally likely bins determined from climatology. Other reasonable choices for the category definitions are pos-
sible, including bins with fixed width or bins that correspond with critical, user-dependent thresholds. However, such additional configurations are not considered in the present study. We choose to use the mode population normalized by the ensemble size (or mode frequency, MOD), as defined in Table 1, as one categorical measure of ensemble spread. Popular categorical error metrics include the ranked probability score (RPS) and the Brier score (BS), so each is included in the sensitivity analysis where appropriate.

A comparison of different stochastic ensemble SD-error correlations reveals that the standard SD-AEM combination is far from the best available choice (Fig. 2). SD is more highly correlated with the non-standard AAE and SDE metrics. It is not surprising that such measures correlate better with ensemble spread than does AEM, because ensemble spread is inherently incorporated into the AAE and SDE error measures. These error measures better reflect the overall forecast ensemble quality by averaging the absolute or squared error of each ensemble member. In addition, it appears that the more similar the calculation between the spread and error metric, the better the linear fit. SD and SDE have nearly identical formulations, since SDE is just the standard deviation of the forecast ensemble about the verification. Although SD-SDE correlations are the largest among any combination explored, the SD-SDE correlations never approach unity, even for large \( \beta \) values. For typical levels of \( \beta \) observed from real-world forecast ensembles (\( \beta = 0.3–0.5 \)), SD-SDE correlations never reach above 0.8. The CRPS, a more widely accepted measure of forecast error based on the full ensemble PDF, surprisingly has a smaller correlation with SD. The SD-CRPS correlation is of similar magnitude to the SD-AEM correlation.

Scatter diagrams for stochastic ensemble spread-error data using SD and three selected measures of error (Fig. 3, top row) show that the joint distributions are not well-described by a simple linear relationship. In general, the diagrams depict increasing forecast error scatter for increasing ensemble spread. The SD-SDE scatter diagram displays the least increase in forecast error scatter with increasing spread and thus possesses the largest linear correlation (0.834). In the case of the AEM metric, because the order of operations is such that ensemble-averaging takes place before the absolute value operator, sign cancellations allow for near-zero AEM even when SD is large. Thus, the lower bound of AEM is always zero for any value
of SD. By definition, the SD-SDE correlations must be larger because the squaring operation happens before the ensemble-averaging step in the SDE calculation, thereby avoiding any error cancellation. The quasi-linear upper error bounds in the SD-AEM scatter diagram explain why Saetra and Bidlot (2004) achieved good results by correlating ensemble spread with the 90th percentile of forecast errors.

Categorical metrics tend to degrade the linear correlation between stochastic ensemble spread and error. Scatter diagrams for the ENT-AEM and ENT-SDE combinations (Fig. 3, bottom row, left and middle panels) exhibit more forecast error scatter and thus lower associated correlation coefficients than for their SD-AEM and SD-SDE counterparts. The sign cancellations associated with the AEM metric remain evident, since low AEM occurs often, even when ENT is relatively large. The lower bound of SDE is not as well defined in the ENT-SDE scatter diagram. In both ENT-AEM and ENT-SDE scatter diagrams, the upper error bounds are fuzzy.

When ENT is correlated with a categorical measure of error like RPS (Fig. 3, bottom row, right panel), an interesting nonlinear pattern emerges. Despite the nonlinear effect, many key features noted in the other scatter diagrams persist. There are clear lower RPS boundaries and the forecast error scatter tends to increase with larger ENT. The primary difference is the bimodal distribution of RPS for low ENT values. Although the vast majority of occurrences occupy the lower branch, there is some non-zero probability of large RPS, even when ENT is small. This is unlike the characteristics of the scatter diagrams for continuous measures of ensemble spread and error, where small ensemble spread constrains the forecast errors to be low. With small ENT values, the forecast is either successful (low RPS, more common for an ideal ensemble) or unsuccessful (high RPS, less common) and moderate RPS values do not fall within the plausible ENT-RPS state space. Moderate RPS values only become attainable when the categorical ensemble spread, as measured by ENT, increases to larger levels.
3 A Probabilistic Approach

As demonstrated from scatter diagrams in Section 2.2, a linear model oversimplifies actual ensemble spread-error relationships. Even for an ideal ensemble prediction system, spread and error are not perfectly correlated using any combination of ensemble size, ensemble-spread temporal variability, or spread and error metric. In reality, a range of forecast errors occur for each value of ensemble spread and how this distribution of forecast errors changes with respect to the ensemble spread is of interest.

Although ensemble-mean forecast errors are generally constrained to be small for low ensemble spread cases, the opposite is not observed for large spread cases (Barker 1991; Molteni et al. 1996; Whitaker and Loughe 1998). The characteristic trait of ensemble spread-error joint distributions is increasing scatter with increasing spread (e.g., Fig. 3). Therefore, the probability of observing a large forecast error increases and the probability of observing a small forecast error decreases (but remains substantial) with increasing ensemble spread. A completely probabilistic approach to the forecast error prediction problem seems appropriate, as suggested by Ziehmann (2001).

A probabilistic approach requires that two principles be fully understood: (1) the distinction between forecast error and forecast uncertainty, and (2) the concept of statistical consistency. These principles are discussed first and then a procedure for measuring the ensemble spread-error relationship with forecast error probability skill scores is described. In the final part of this section, the results of the error probability forecast analysis are presented using stochastic ensemble data.

3.1 The distinction between error and uncertainty

It is usually straightforward to define forecast error in the deterministic context by a simple difference between an observation and a forecast. Yet, how should this discrepancy be measured when the forecast is expressed as a PDF and the observation is only one discrete value? If one possesses an ideal ensemble prediction system, in which all sources of forecast error are properly characterized, then the forecast PDF from such a system is an expression of the actual
forecast uncertainty. Of course, ideal NWP ensembles do not really exist, so the actual forecast uncertainty (or true PDF) is normally unknown and cannot be estimated for an individual case (de Elía and Laprise 2005). Therefore, the quality of any individual forecast PDF is not well defined. Only the aggregate quality of a collection of forecast PDFs can be well estimated, using realizations from several cases (Wilks 2005).

Provided that forecast error is measured by the difference of the ensemble-mean from the observation in each case, having a collection of these errors yields information about the actual forecast uncertainty averaged over the sample. This concept is illustrated by Fig. 4, which shows ensemble forecast PDFs (solid curves) and corresponding observations (thick gray vertical lines) for the first three cases (top three rows) of a 400-case sample using the stochastic model of Section 2.1. To readily distinguish the stochastic ensemble PDF curves from the ideal forecast PDF curves (dashed), the stochastic ensemble is modified here to be underdispersed. This is accomplished by prescribing the stochastic ensemble to have 25% of the ideal variance\(^5\) for each case. Having deficient ensemble spread also simulates the typical performance of real ensemble prediction systems. The simulated observations are random draws from the ideal PDFs for each case. Ensemble-mean forecast errors are recorded in a histogram that evolves as more cases are added (right column). After 400 cases (bottom right), it is obvious that the ensemble-mean forecast error distribution corresponds well with the actual forecast uncertainty averaged over that sample (bottom left, dashed curve), even though the stochastic ensemble is imperfect in this example. This condition holds for any forecast ensemble because ensemble-mean forecast errors can be considered random draws from each ideal PDF, just like the observations. Therefore, the distinction between error and uncertainty is that an individual error is merely one sample from a distribution that represents the uncertainty. A distribution of ensemble-mean forecast errors collected from many cases indicates the average actual forecast uncertainty over that sample. This characteristic suggests that forecast ensemble PDFs ought to be compared with the distributions of ensemble-mean forecast errors, not with single values of ensemble-mean forecast error.

\(^5\)Recall that, the actual forecast uncertainty is known in these simple numerical experiments.
3.2 Statistical consistency

When forecast ensembles and ensemble-mean error distributions match up well, the ensemble prediction system exhibits statistical consistency (Anderson 1996; Talagrand et al. 1997). A statistically consistent ensemble prediction system thus produces reliable (calibrated) forecast probabilities at all thresholds. Statistical consistency is often measured by the uniformity of verification rank histograms (Anderson 1996; Hamill 2001), since verifications ought to fall equally often among the bins defined by the ordered ensemble member forecasts. Statistical consistency can also be evaluated in a broader sense by noting how closely the average ensemble variance matches up with the ensemble-mean error variance (Talagrand et al. 1997; Ziehmann 2000; Eckel and Mass 2005).

A key point here is that statistical consistency is usually evaluated over a large sample, preferably with data that are uncorrelated in time and space (Hamill 2001). Therefore, statistical consistency is nominally thought of as a condition fulfilled over many independent samples. Nothing precludes the assessment of statistical consistency for subsets of cases, as long as the sample size remains large and the spatio-temporal correlation in the data remains negligible. In fact, a forecast ensemble that possesses statistical consistency over separate sub-samples (e.g., those cases with low ensemble spread), in addition to the aggregate sample, passes a more stringent test of quality.

Ensemble spread-error analyses really just provide stringent tests of ensemble quality, since they evaluate the statistical consistency of an ensemble prediction system for each ensemble spread class. Wang and Bishop (2003) point out this fact and analyze the performance of an ensemble data assimilation system over separate classes of ensemble spread using comparisons of ensemble variance and ensemble-mean error variance. An ideal ensemble prediction system, which exhibits statistical consistency, should also have a perfect ensemble-variance/error-variance relationship. The stochastic ensemble spread-error scatter diagrams in Fig. 5 emphasize this point. Fig. 5a shows the SD-AEM scatter and the associated least-squares regression lines for 50-member stochastic ensembles with increasing ensemble-spread temporal variability ($\beta = 0.1, 0.3$, and $0.5$ from left to right). The SD-AEM linear correlations range from
about 0.12 to 0.51 and give no obvious indication that the stochastic ensemble system used to generate this data is an ideal one. Using the same ideal ensemble spread-error data, but retaining the sign of the ensemble-mean errors, one observes a nearly perfect linear correlation between the ensemble spread and the spread of the ensemble-mean errors (open circles, Fig. 5b). Statistical sampling is the only factor that prevents seeing perfect correlations. The inherent statistical consistency of the ideal ensemble system is now evident from this one-to-one association between ensemble variance and ensemble-mean error variance. Even in the scenario with low ensemble-spread variability ($\beta = 0.1$), ensemble spread correlates well with the spread of the ensemble-mean errors (Fig. 5b, left panel). As $\beta$ increases, the resolved range of ensemble-mean error variances increases, allowing for improved discrimination between cases with small or large forecast uncertainty. Thus, the usefulness of ensemble spread as a predictor of forecast error is predicated on its ability to distinguish between different forecast error distributions. An ideal ensemble prediction system exhibits a perfect ensemble-variance/error-variance relationship. Therefore, one ought use a tool that objectively measures this facet.

### 3.3 Conditional-error-climatology (CEC) predictions based on ensemble spread

We propose a tool which measures the ensemble-variance/error-variance relationship by the impact that ensemble spread has on improving error probability forecasts compared to climatology. Within the ideal ensemble scenario studied here, it makes little sense to use any error probability forecast other than the one directly generated by the stochastic ensemble. An error probability forecast based on an ideal ensemble PDF is the best one can make, due to the lack of bias and the perfect case-to-case resolution of the actual forecast uncertainty. However, non-ideal forecast ensembles are likely to contain prominent biases in the first two moments (mean and variance). This motivates the use of historical forecast error statistics to correct such biases. For example, one simple approach is to “dress” the ensemble-mean forecast with its historical error distribution (Atger 1999). Probability forecasts based on this error climatology
possesses no skill, since they are temporally invariant and based on some long-term average forecast error condition. A subset of the error climatology, or conditional-error climatology (CEC), has the potential to perform better as an error probability forecast if anomalous forecast uncertainties are shared by the chosen subset of cases. If CEC is constructed by using histories of ensemble-mean forecast errors associated with specific classes of ensemble spread, then any improvement in error probability forecast skill, relative to the error climatology forecast performance, can be directly attributed to the ensemble-variance/error-variance relationship.

Construction of spread-based CEC forecasts really amounts to a statistical post-processing technique, where ensemble-mean forecasts are “dressed” with the forecast error distribution from historical cases with similar ensemble spread. Therefore, these CEC forecasts are like simple analog forecasts (Wilks 2005), where the statistical analogs are chosen in a univariate sense rather than from multivariate forecast fields. For the stochastic ensemble data, CEC forecasts are constructed by separating the $N$ idealized cases into groups according to their corresponding ensemble variance. This is the same grouping technique illustrated in Fig. 5b, where ensemble-mean forecast errors are binned into 20 variance categories. The chosen number of ensemble-variance categories is arbitrary. Clearly, more subdivisions provide the potential for higher skill, at least within the groups containing forecast error distributions most distinct from the overall error climatology. The choice is limited by the sample size of the data set. To ensure that high statistical significance is reached at all times, each category is required to have at least $10^3$ samples in this experiment, resulting in the use of a maximum of nine ensemble-variance categories.

Ensemble spread-based CEC forecasts are objectively compared with two reference uncertainty forecasts. The ensemble-mean error climatology (ERR-CLI) serves as a baseline (no-skill) probability forecast. The raw ensemble plays the role of a target (perfect-skill) probability forecast, since the stochastic ensemble produces ideal forecasts of uncertainty. To evaluate the improvement of spread-based CEC forecasts over ERR-CLI forecasts using the stochastic ensemble data, a cross-validation procedure is used. Cycling through the $N$ cases, each case is individually withheld as test data and the remaining $N - 1$ cases are treated as training data. Error climatology is not necessarily thought of as a historical record of forecast
errors in this context, since chronology is not important in the stochastic ensemble data and all $N - 1$ training cases are available for the analog selection at the time of processing. In practice, only historical data could be used to estimate the error climatologies.

Ensemble spread-based CEC forecasts are developed for both continuous (VAR-CEC) and categorical (ENT-CEC and MOD-CEC) measures of ensemble spread and are tested under different end-user specifications. For end users who have continuous sensitivity to forecast error, spread-based CEC forecasts are smoothed by fitting a Gaussian distribution to the errors in each spread bin using the sample mean and variance of the errors. This produces a continuous forecast PDF. To facilitate a fair comparison, a Gaussian fit is also applied to the raw forecast ensemble (ENS-FIT) using the ensemble mean and variance. These continuous error probability forecasts are evaluated using the CRPS, a measure that is minimized for probability forecasts that are statistically consistent and have maximum sharpness (Gneiting et al. 2005). CRPS can be interpreted as the probabilistic analog of mean absolute error for deterministic forecasts.

Statistical consistency of the error probability forecasts is evaluated by the uniformity of probability integral transform (PIT) histograms, which are the analogs of verification rank histograms for continuous forecast PDFs (Gneiting et al. 2005). The PIT value is defined as

$$ PIT = \int_{-\infty}^{Y} p(x) dx, $$

where $p(x)$ is the forecast probability density function and $Y$ is the verification. The PIT value is just the forecast cumulative distribution function evaluated at the verification, which is the area under the forecast PDF curve to the left of the verifying value. Histograms of PIT values ought to be uniform for the same reasons that verification rank histograms should be uniform.

The Ziehmann (2001) methodology is used to determine CEC for a categorical end user. Specifically, the forecasts and verifications are divided into the same climatological bins used to determine ENT or MOD and the $N - 1$ training cases are partitioned into categories according to the magnitude of the ensemble spread metric for each case. The ensemble spread for the given test case is calculated and its appropriate category determined. The categorical CEC is determined from the forecast errors of all of the training cases falling within the same spread
category as the test case. For a binary forecast verification framework using an error measure like success/failure, the rate of success\textsuperscript{6} is determined for the cases in the given ensemble spread category. This quantity is called the *conditional success rate* and represents the probability of success for cases with ensemble spread similar to the test case. Hence, the conditional success rates are used as probability of success forecasts since they are the categorical analogs of CEC PDFs for continuous data. Evaluation of the quality of the categorical probability forecasts is accomplished using the Brier score (BS).

### 3.4 Assessment of the error probability forecasts

The predictive skill of the various error probability forecasts using the stochastic ensemble data is first evaluated from the perspective of an end user who has a continuous sensitivity to forecast error (Fig. 6). Three different measures of ensemble spread are used to form the conditional-error-climatology predictions from a 50-member stochastic ensemble: one continuous (VAR-CEC) and two categorical (ENT-CEC and MOD-CEC). The predictive skill of each error probability forecast is presented as a function of the stochastic ensemble-variance category and measured with CRPS and its associated skill score (CRPSS) relative to the ensemble-mean error climatology (ERR-CLI) forecast\textsuperscript{7}. In general, the CRPS increases (skill decreases) as the stochastic ensemble variance grows for all forecast types, which is simply a reflection of the decreasing sharpness in the forecasts. The skill scores are distributed over the stochastic ensemble-variance categories with an asymmetric u-shape that favors the low-variance categories over the high ones. The u-shaped CRPSS distributions reinforce the findings of earlier studies (Houtekamer 1993; Whitaker and Loughe 1998; Grimit and Mass 2002), showing that forecast error predictability is greater for cases with anomalous (high or low) spread than for cases with near-normal spread. Because the ensemble-mean errors for near-normal spread cases are distributed about the same as the overall ensemble-mean error climatology, there is little advantage to be gained by using them.

\textsuperscript{6}The success rate is defined as the fraction of cases where the ensemble mean falls into the same climatological bin as the verification.

\textsuperscript{7}CRPSS = 1 – CRPS/CRPS\textsubscript{ERR-CLI}
Given that the stochastic ensemble forecasts are ideal, the Gaussian-fitted ensemble (ENS-FIT) forecasts ought to verify the best. Comparisons between the CEC methods and ENS-FIT forecasts indeed show that ENS-FIT is more skillful for every variance category. However, the aggregated CRPSS for the entire $10^4$-case sample (Table 3) using nine spread categories is 0.06 for ENS-FIT and 0.055 for VAR-CEC. Thus, VAR-CEC forecasts have only slightly less overall skill than ENS-FIT forecasts. ENT-CEC performs slightly better than MOD-CEC, which is a reflection of the fact that ENT is more consistent with the continuous verification framework since it is calculated using the entire forecast distribution, whereas MOD is not. Using fewer ensemble-variance categories systematically reduces the skill of the spread-based CEC methods, independent of the particular spread metric. In parallel with the spread-error correlation results, the predictive skill of all error probability forecast types is reduced as the ensemble size $M$ and the ensemble-spread temporal variability $\beta$ are decreased (not shown).

The relative performance of the methods as well as the u-shape of the skill score distributions as a function of the ensemble-variance categories can be understood by examining probability integral transform (PIT) histograms for the error probability forecasts. The PIT histograms (Fig. 7) show over-dispersion (peaked histograms) for ERR-CLI forecasts for cases with anomalous ensemble spread, while the PIT histograms for VAR-CEC and ENS-FIT forecasts are essentially uniform over all variance categories. These characteristics lead to the u-shape in the CRPSS distributions, since VAR-CEC and ENS-FIT forecasts make their improvements over ERR-CLI forecast exclusively in anomalous spread situations. ENT-CEC and MOD-CEC forecasts produce distributions of PIT histograms similar to ERR-CLI forecast PIT histograms, but are somewhat closer to uniformity (statistical consistency) for each variance category. The PIT histograms for ENT-CEC and MOD-CEC are not uniform because ENT and MOD do not distinguish between cases with small and large uncertainty as well as ensemble variance does for the continuous verification scenario.

The asymmetry of the CRPSS distributions favoring the low ensemble-variance categories can be explained by the effect of sharpness. Since the ERR-CLI probability forecasts are constant predictions of forecast uncertainty, the sharpness (measured here by the average width of the 90% central prediction intervals) of these forecasts is also constant (Fig. 8). The
sharpness for ENS-FIT and VAR-CEC forecasts is better (worse) than for ERR-CLI forecasts in cases with small (large) ensemble spread. Because sharper forecasts generate a lower CRPS, this leads to the asymmetric skew. The ENS-FIT forecasts are slightly sharper than VAR-CEC forecasts overall and this leads to the slight advantage in skill.

The relative performance of the CEC methods using the various spread metrics is quite different for end users who do not have a continuous cost function. After dividing the stochastic ensemble forecasts and verifications into ten equally-likely climatological bins, a probability of success is computed for each CEC method and ENS-FIT. For the CEC methods, this conditional success rate is based on the fraction of cases with similar spread values in which the ensemble mean falls into the same climatological bin as the verification. The probability of success from ENS-FIT is given by the forecast probability density within the climatological bin containing the ensemble mean. The BS and its associated skill score (BSS) calculated relative to the climatological success rate are reported for ENS-FIT and each CEC method (Fig. 9 and Table 4). As before, the skill score distributions take on an asymmetric u-shape, slanted in favor of the low spread events. Unlike the continuous case, the skill scores do not reach zero for cases with near-normal spread. This means that probability of success forecasts for near-normal spread cases have value. In addition, categorical CEC forecasts now outperform VAR-CEC forecasts, with ENT-CEC and MOD-CEC forecasts scoring only slightly worse than ENS-FIT forecasts. Therefore, categorical (continuous) measures of forecast ensemble spread are more appropriate for end users with a categorical (continuous) sensitivity to forecast error. The difference in relative performance between the two verification methods is further evidence that the chosen metric for spread should be consistent with the cost function of the target end user.

4 Conclusions

In this study, ideal ensemble spread and error statistics are generated by a simple stochastic model for varying ensemble sizes and ensemble-spread temporal variances. The stochastic ensemble spread and error statistics exhibit a relationship that is characterized by increasing
error variance with growing ensemble spread using a variety of spread and error metrics. Such a relationship is not well described by a linear least-squares fit between the spread and error, since it does not account for the varying uncertainty as a function of ensemble spread. Thus, spread-error linear correlations oversimplify the true joint distribution of ensemble spread and error. Furthermore, error forecasts based on a linear regression equation implicitly assume that the target end user has a continuous sensitivity to forecast error. While some end users clearly have cost structures that are continuous functions of forecast error, a large percentage of end users have a categorical sensitivity to forecast error that often depends on critical thresholds, such as freezing temperatures. For categorical situations, alternate metrics for ensemble spread and forecast error are necessary and linear correlations between these measures are inappropriate. It is suggested that the ensemble spread-error relationship is better measured with a probabilistic approach, since the distribution of forecast errors changes as the ensemble spread varies, with the probability of large (small) forecast errors increasing (decreasing) as ensemble spread grows larger.

A probabilistic approach to measuring the ensemble spread-error relationship is explored with a large sample ($10^4$ cases) of stochastic ensemble data. The new approach recognizes the inherent nonlinearity of ensemble spread-error joint distributions and can be easily adapted to handle a broad range of end-user sensitivity to forecast error. It measures the connection between ensemble variance and ensemble-mean error variance, since statistically consistent ensemble prediction systems should exhibit a perfect ensemble-variance/error-variance relationship. In particular, the ensemble-mean errors from cases with similar ensemble spread as a given test case are used as a probability forecast for the ensemble-mean error for that test case. These error probability forecasts are compared with reference uncertainty forecasts based on the overall, unconditional error climatology. The resulting skill score functions as a practical measure of the ensemble-variance/error-variance relationship. That is, the ultimate benefit realized by end users is quantified by the verifying skill of the spread-based error probability forecasts compared to climatology.

Conditional-error-climatologies may be produced using any reasonable metric of ensemble spread, such as a continuous measure like standard deviation or a categorical measure like
mode population. The skill of conditional-error-climatology forecasts based on ensemble spread depends on the number of ensemble-spread categories used. As the number of categories is increased, the skill scores improve. With nine ensemble spread categories that are derived from a 50-member, stochastic forecast ensemble characterized by moderate ensemble spread temporal variability ($\beta = 0.5$), conditional-error-climatology forecasts based on ensemble variance produce a continuous ranked probability skill score of 5.5%. For categorical forecasts using ten climatological bins and the same stochastic ensemble data, the probability forecasts based on the ensemble mode frequency produce a Brier skill score of 17.2%. Smaller ensemble sizes and lower ensemble-spread temporal variabilities result in lower skill scores. These improvements in bulk scores are modest, implying that even for an ideal ensemble prediction system, there is not a great deal of predictive skill to be gained by using the ensemble spread. However, the actual benefit realized by specific end users of these forecasts could be very wide-ranging and heavily dependent on the cost function involved.

The predictive skill for conditional-error-climatology forecasts based on stochastic ensemble spread is greater for cases with anomalous ensemble spread than for cases with near-normal spread, validating conclusions from earlier empirical studies. Additionally, the greatest predictive skill exists for small spread cases because the error probability forecasts are inherently sharper. Therefore, ensemble spread is most useful as a predictor of error probability mainly when it is anomalous and especially when it is low. In the smallest of nine ensemble-variance categories that are derived from a 50-member stochastic ensemble characterized by moderate ensemble-spread variability ($\beta = 0.5$), skill scores reach up to 30–40%. The relative frequency of anomalous spread occurrence, which is quantified by the ensemble-spread variability, ultimately controls how much valuable information can be gleaned from the ensemble spread. The available evidence from existing ensemble NWP systems points towards a situation in which the ensemble spread remains relatively constant from one day to the next, at least compared with the ensemble-mean fluctuations (Jewson 2004). If this lack of ensemble-spread variability extends to most weather forecast parameters of interest to end users, then perhaps the main efforts of forecast improvement should be directed toward reducing ensemble-mean errors (e.g., via improvements to model physical parameterizations) rather than toward quantifying the
state-dependent forecast uncertainty about that ensemble mean. Constant uncertainty forecasts taken from the ensemble-mean error climatology might suffice (Atger 1999), with so little to be gained from the ensemble spread-error relationship in such a scenario.

The ensemble spread metric that produces the most skillful error probability forecast depends on the verification framework. With the stochastic ensemble data, the conditional-error-climatology that performs best for continuous forecasts is the one determined by the ensemble variance, which is a continuous measure of spread. For categorical situations, conditional-error-climatology forecasts based on the ensemble mode population work well. Therefore, continuous (categorical) measures of ensemble spread are most appropriate for end users with a continuous (categorical) cost function. By extension, ensemble spread ought to be quantified by a metric that is as closely related to a target end users’ cost function as possible. This preference for metric consistency is also evident from the spread-error correlation sensitivity results. The need for metric consistency at first appears to conflict with the results of Toth et al. (2001) and Ziehmann (2001), who find that categorical spread metrics produce superior error forecasts. However, those studies rely solely on a categorical verification framework and therefore do not consider end users with continuous cost functions. The results of this study support the use of indices like the NCEP relative measure of predictability, if the target end user has a categorical cost function that is aligned with the same climatological bins used to define the ensemble spread metric.

To guarantee good results with categorical spread metrics, the ensemble size must always be large compared to the number of climatological bins that are chosen. With smaller ensemble sizes, there exist fewer combinations of values that statistical entropy and mode frequency can assume, thereby decreasing the ability to discriminate between successes and failures. For the method to remain viable using small ensemble sizes, a reduced number of climatological bins would be necessary. With 10 equally-likely climatological bins and ensemble sizes less than 50, we encountered problems with the partitioning of forecast errors using categorical spread metrics.

Conditional-error-climatology forecasts based on stochastic ensemble spread perform quite well compared to probability forecasts based directly on the stochastic ensemble sample statis-
tics. There is less than a 1% difference in the overall skill scores between the two methods when nine ensemble-spread categories are used. This suggests that, in addition to being a good tool for measuring the real NWP ensemble spread-error relationships, spread-based conditional-error-climatology forecasts could serve as a viable post-processing methodology for such systems. These calibrated probability forecasts have the potential to significantly outperform raw ensemble output in a real-world implementations.

The main impediment to the creation of error climatologies is a lack of sample size. Real NWP ensemble systems are frequently upgraded, making the definition of climatology an evolving target. Having to further divide error climatologies by ensemble spread classes only exacerbates the problem. Another disadvantage is that conditional-error-climatologies are generally unimodal (and Gaussian in this case). To the extent that multi-modality is appropriate and short training periods are necessary, a post-processing method such as Bayesian model averaging (Raftery et al. 2005) presents several advantages. Provided that a very large historical archive of forecast ensemble data is available, error climatologies could be further split according to the forecast state, not just the ensemble spread. Analog selection of this type appears to have considerable potential for probability forecast calibration (e.g., Roulston et al. 2003; Hamill et al. 2006).

In the future, the need for ensemble spread-error analysis will be reduced as better post-processing methods are developed and the ensemble prediction systems themselves are improved. The ensemble spread-error relationship may only be needed as a tool for checking that statistical consistency is achieved over all ensemble spread classes. At present, the use of ensemble spread-error relationships as a competitive method of post-processing remains viable, especially since the weather forecast community still heavily relies upon deterministic predictions from an ensemble mean or high-resolution control run. Having a reasonable estimate of the uncertainty of these deterministic forecasts is valuable, even if it is produced from a suboptimal ensemble prediction system that is underdispersive and run at lower resolution.
Acknowledgments

This work represents a portion of the first author’s Ph.D. dissertation under the supervision of the second author at the University of Washington. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) under Office of Naval Research Grant N00014-01-10745. Tilmann Gneiting and two anonymous reviewers greatly enhanced the final version of this manuscript.
References


## List of Tables

1. Acronyms and definitions for the various measures of forecast ensemble spread used in this paper, where $M$ is the ensemble size, $x_m$ refers to forecast ensemble member $m$, $\overline{x}$ denotes the ensemble mean, $nbin$ is the number of categories defined from climatology (set to ten in this paper), and $M_i$ is the population of forecast ensemble members within climatological bin $i$. ........................................ 35

2. Acronyms and definitions for the various measures of forecast error used in this paper, where $M$ is the ensemble size, $x_m$ refers to forecast ensemble member $m$, $\overline{x}$ denotes the ensemble mean, $y$ is the verification, $s$ is the ensemble standard deviation (SD), $\delta_{y \in C_i}$ represents a binary verification equal to 1 if the verification falls into climatological bin $i$ ($C_i$) and equal to 0 if it does not, $p_i$ is the probability density contained within $C_i$, and $p_{\overline{x}}$ is the probability density contained within the climatological bin containing $\overline{x}$ ($C_{\overline{x}}$). The expression given for CRPS assumes a Gaussian forecast probability density function with mean $\overline{x}$ and standard deviation $s$, where $\Phi(x)$ represents the standard Gaussian cumulative distribution function evaluated at $x$, and $\phi(x)$ is the standard Gaussian probability density function evaluated at $x$ (see Gneiting et al. 2005). ........... 36

3. Continuous ranked probability skill score (CRPSS) over the entire $10^4$-case sample of 50-member stochastic ensemble data with $\beta = 0.5$ for the ENS-FIT, VAR-CEC, ENT-CEC, and MOD-CEC error probability forecasts as a function of the number of forecast ensemble-variance categories. ........................................ 37

4. Brier skill score (BSS) over the entire $10^4$-case sample of 50-member stochastic ensemble data with $\beta = 0.5$ for the ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC probability of success forecasts as a function of the number of forecast ensemble-variance categories. ........................................ 38
List of Figures

1 Spread-error correlation coefficients for the stochastic ensemble outlined in Section 2.1 using the ensemble standard deviation (SD) as a measure of spread and the absolute error of the ensemble mean (AEM) as a measure of error. Stochastic ensemble size (M) varies from 2–50 and ensemble-spread temporal variability ($\beta$) ranges from 0.1–0.9. .......................... 39

2 Stochastic ensemble spread-error correlation sensitivity to the error metric, assuming $\beta = 0.5$. Ensemble standard deviation (SD) is used as a measure of spread. Measures of error included here are the absolute error of the ensemble mean (AEM), the average absolute error over the ensemble (AAE), the standard deviation of the error over the ensemble (SDE), and the continuous ranked probability score (CRPS). .......................... 40

3 Selected ensemble spread-error scatter diagrams using a continuous measure of ensemble spread (SD, top row) and a categorical measure of ensemble spread (ENT, bottom row) for a stochastic ensemble with $M = 50$ and $\beta = 0.5$. Note that, ENT-AEM and ENT-SDE correlations pair a categorical measure of spread with a continuous measure of forecast error. ENT-RPS correlations are composed of two categorical measures. Only the first 3000 cases are plotted to reduce clutter. The empirical linear correlations are reported in each panel. 41
A schematic illustrating how an imperfect, stochastic ensemble produces an ensemble-mean error distribution that approximates the average actual forecast uncertainty over the cases making up the sample. The left column depicts the first three individual cases (top three panels) of a 400-case sample of 50-member, stochastic ensemble forecasts prescribed to have deficient variance (solid curves) as well as observations for each case (thick gray vertical lines). Also indicated are the PDFs representing the actual forecast uncertainty for each case (dashed curves, normally unknown) from which the observations are randomly simulated. The average ensemble forecast PDF and the average actual forecast uncertainty over the 400-case sample are shown in the bottom panel in a reference frame relative to the ensemble mean (mean-removed) so they may be easily compared. The right column displays the corresponding ensemble-mean forecast error histogram as it evolves through the cases. After all 400 cases (bottom panel), the forecast error histogram well approximates the average actual forecast uncertainty.

Ensemble spread-error scatter diagrams for 50-member stochastic ensembles with $\beta = 0.1$, 0.3, and 0.5 from left to right. (a) The SD-AEM scatter with linear least-squares regression lines (dashed) and correlation coefficients reported in each panel. (b) The same data as in (a), except retaining the sign of the ensemble-mean forecast errors. Overlaid are the ensemble-mean forecast error standard deviations (open circles) calculated for 20 ensemble-spread classes. The linear association between the ensemble spread and the ensemble-mean error spread is nearly perfect, with the open circles lying very near the diagonal, dashed line indicating statistical consistency.
Error probability forecast skill as a function of nine ensemble-variance categories using 50-member stochastic ensemble data with $\beta = 0.5$. The Gaussian-fitted ensemble forecasts (ENS-FIT) and the ensemble variance-based, statistical entropy-based, and modal frequency-based conditional-error-climatology forecasts (VAR-CEC, MOD-CEC, and ENT-CEC) are scored with the continuous ranked probability score (CRPS). The associated skill score (CRPSS) is calculated relative to the CRPS for the ensemble-mean error climatology (ERR-CLI) forecast.

Probability integral transform (PIT) histograms (the analog of verification rank histograms for continuous forecast PDFs) as a function of nine ensemble variance categories (columns 1–9) for ERR-CLI, ENS-FIT, VAR-CEC, ENT-CEC, and MOD-CEC probabilistic error forecasts. PIT histograms for the total $10^4$-case sample are indicated in the far right column.

Probabilistic forecast sharpness using average 90% central prediction interval widths as a function of nine ensemble-variance categories for ERR-CLI, ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC error probability forecasts. The average central prediction interval widths are normalized by the climatological median ensemble standard deviation, $\overline{\sigma}_F$.

Predictive skill as a function of nine ensemble-variance categories for ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC probability of success forecasts measured by the Brier skill score (BSS).
Table 1: Acronyms and definitions for the various measures of forecast ensemble spread used in this paper, where $M$ is the ensemble size, $x_m$ refers to forecast ensemble member $m$, $\bar{x}$ denotes the ensemble mean, $nbin$ is the number of categories defined from climatology (set to ten in this paper), and $M_i$ is the population of forecast ensemble members within climatological bin $i$.

<table>
<thead>
<tr>
<th>Spread Metric</th>
<th>Acronym</th>
<th>Mathematical Expression</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard deviation</td>
<td>SD</td>
<td>$\sqrt{\frac{1}{M-1} \sum_{m=1}^{M} (x_m - \bar{x})}$</td>
<td>continuous</td>
</tr>
<tr>
<td>variance</td>
<td>VAR</td>
<td>$\frac{1}{M-1} \sum_{m=1}^{M} (x_m - \bar{x})$</td>
<td>continuous</td>
</tr>
<tr>
<td>mode frequency</td>
<td>MOD</td>
<td>$\frac{1}{M} \max(M_i)$</td>
<td>categorical</td>
</tr>
<tr>
<td>statistical entropy</td>
<td>ENT</td>
<td>$- \sum_{i=1}^{nbin} \frac{M_i}{M} \log \left( \frac{M_i}{M} \right)$</td>
<td>categorical</td>
</tr>
</tbody>
</table>
Table 2: Acronyms and definitions for the various measures of forecast error used in this paper, where $M$ is the ensemble size, $x_m$ refers to forecast ensemble member $m$, $\bar{x}$ denotes the ensemble mean, $y$ is the verification, $s$ is the ensemble standard deviation (SD), $\delta_{y \in C_i}$ represents a binary verification equal to 1 if the verification falls into climatological bin $i$ ($C_i$) and equal to 0 if it does not, $p_i$ is the probability density contained within $C_i$, and $p_{\bar{x}}$ is the probability density contained within the climatological bin containing $\bar{x}$ ($C_{\bar{x}}$). The expression given for CRPS assumes a Gaussian forecast probability density function with mean $\bar{x}$ and standard deviation $s$, where $\Phi(x)$ represents the standard Gaussian cumulative distribution function evaluated at $x$, and $\phi(x)$ is the standard Gaussian probability density function evaluated at $x$ (see Gneiting et al. 2005).

<table>
<thead>
<tr>
<th>Error Metric</th>
<th>Acronym</th>
<th>Mathematical Expression</th>
<th>Cost Function Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolute error of the ensemble mean</td>
<td>AEM</td>
<td>$</td>
<td>\bar{x} - y</td>
</tr>
<tr>
<td>average absolute error (over the ensemble)</td>
<td>AAE</td>
<td>$\frac{1}{M} \sum_{m=1}^{M}</td>
<td>x_m - y</td>
</tr>
<tr>
<td>standard deviation of the error (over the ensemble)</td>
<td>SDE</td>
<td>$\sqrt{\frac{1}{M-1} \sum_{m=1}^{M} (x_m - \bar{x})^2}$</td>
<td>continuous, linear</td>
</tr>
<tr>
<td>continuous ranked probability score</td>
<td>CRPS</td>
<td>$s \left( \frac{\Phi\left( \frac{y - \bar{x}}{s} \right)}{\Phi\left( \frac{y - \bar{x}}{s} \right) - 1} + 2\phi\left( \frac{y - \bar{x}}{s} \right) - \frac{\Phi\left( \frac{y - \bar{x}}{s} \right)}{\Phi\left( \frac{y - \bar{x}}{s} \right) - 1} \right)$</td>
<td>continuous, linear</td>
</tr>
<tr>
<td>failure (1 - success)</td>
<td>n/a</td>
<td>$1 - \delta_{y \in C_i}$</td>
<td>categorical, binary</td>
</tr>
<tr>
<td>Brier score</td>
<td>BS</td>
<td>$(p_{\bar{x}} - \delta_{y \in C_{\bar{x}}})^2$</td>
<td>categorical, binary</td>
</tr>
<tr>
<td>ranked probability score</td>
<td>RPS</td>
<td>$\sum_{n=1}^{n_{bin}} \left( (\sum_{i=1}^{n} p_i) - \left( \sum_{i=1}^{n} \delta_{y \in C_i} \right)^2 \right)$</td>
<td>categorical, multiple</td>
</tr>
</tbody>
</table>
Table 3: Continuous ranked probability skill score (CRPSS) over the entire $10^4$-case sample of 50-member stochastic ensemble data with $\beta = 0.5$ for the ENS-FIT, VAR-CEC, ENT-CEC, and MOD-CEC error probability forecasts as a function of the number of forecast ensemble-variance categories.

<table>
<thead>
<tr>
<th>No. Spread Bins</th>
<th>ENS-FIT</th>
<th>VAR-CEC</th>
<th>ENT-CEC</th>
<th>MOD-CEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.055</td>
<td>0.041</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>5</td>
<td>0.058</td>
<td>0.051</td>
<td>0.026</td>
<td>0.018</td>
</tr>
<tr>
<td>7</td>
<td>0.060</td>
<td>0.054</td>
<td>0.027</td>
<td>0.020</td>
</tr>
<tr>
<td>9</td>
<td>0.060</td>
<td>0.055</td>
<td>0.027</td>
<td>0.021</td>
</tr>
</tbody>
</table>
Table 4: Brier skill score (BSS) over the entire $10^4$-case sample of 50-member stochastic ensemble data with $\beta = 0.5$ for the ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC probability of success forecasts as a function of the number of forecast ensemble-variance categories.

<table>
<thead>
<tr>
<th>No. Spread Bins</th>
<th>ENS-FIT</th>
<th>VAR-CEC</th>
<th>ENT-CEC</th>
<th>MOD-CEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.179</td>
<td>0.038</td>
<td>0.145</td>
<td>0.155</td>
</tr>
<tr>
<td>5</td>
<td>0.180</td>
<td>0.051</td>
<td>0.162</td>
<td>0.167</td>
</tr>
<tr>
<td>7</td>
<td>0.179</td>
<td>0.061</td>
<td>0.168</td>
<td>0.171</td>
</tr>
<tr>
<td>9</td>
<td>0.179</td>
<td>0.061</td>
<td>0.171</td>
<td>0.172</td>
</tr>
</tbody>
</table>
Figure 1: Spread-error correlation coefficients for the stochastic ensemble outlined in Section 2.1 using the ensemble standard deviation (SD) as a measure of spread and the absolute error of the ensemble mean (AEM) as a measure of error. Stochastic ensemble size (M) varies from 2–50 and ensemble-spread temporal variability (β) ranges from 0.1–0.9.
Figure 2: Stochastic ensemble spread-error correlation sensitivity to the error metric, assuming $\beta = 0.5$. Ensemble standard deviation (SD) is used as a measure of spread. Measures of error included here are the absolute error of the ensemble mean (AEM), the average absolute error over the ensemble (AAE), the standard deviation of the error over the ensemble (SDE), and the continuous ranked probability score (CRPS).
Figure 3: Selected ensemble spread-error scatter diagrams using a continuous measure of ensemble spread (SD, top row) and a categorical measure of ensemble spread (ENT, bottom row) for a stochastic ensemble with $M = 50$ and $\beta = 0.5$. Note that, ENT-AEM and ENT-SDE correlations pair a categorical measure of spread with a continuous measure of forecast error. ENT-RPS correlations are composed of two categorical measures. Only the first 3000 cases are plotted to reduce clutter. The empirical linear correlations are reported in each panel.
Figure 4: A schematic illustrating how an imperfect, stochastic ensemble produces an ensemble-mean error distribution that approximates the average actual forecast uncertainty over the cases making up the sample. The left column depicts the first three individual cases (top three panels) of a 400-case sample of 50-member, stochastic ensemble forecasts prescribed to have deficient variance (solid curves) as well as observations for each case (thick gray vertical lines). Also indicated are the PDFs representing the actual forecast uncertainty for each case (dashed curves, normally unknown) from which the observations are randomly simulated. The average ensemble forecast PDF and the average actual forecast uncertainty over the 400-case sample are shown in the bottom panel in a reference frame relative to the ensemble mean (mean-removed) so they may be easily compared. The right column displays the corresponding ensemble-mean forecast error histogram as it evolves through the cases. After all 400 cases (bottom panel), the forecast error histogram well approximates the average actual forecast uncertainty.
Figure 5: Ensemble spread-error scatter diagrams for 50-member stochastic ensembles with $\beta = 0.1$, 0.3, and 0.5 from left to right. (a) The SD-AEM scatter with linear least-squares regression lines (dashed) and correlation coefficients reported in each panel. (b) The same data as in (a), except retaining the sign of the ensemble-mean forecast errors. Overlaid are the ensemble-mean forecast error standard deviations (open circles) calculated for 20 ensemble-spread classes. The linear association between the ensemble spread and the ensemble-mean error spread is nearly perfect, with the open circles lying very near the diagonal, dashed line indicating statistical consistency.
Figure 6: Error probability forecast skill as a function of nine ensemble-variance categories using 50-member stochastic ensemble data with $\beta = 0.5$. The Gaussian-fitted ensemble forecasts (ENS-FIT) and the ensemble variance-based, statistical entropy-based, and modal frequency-based conditional-error-climatology forecasts (VAR-CEC, MOD-CEC, and ENT-CEC) are scored with the continuous ranked probability score (CRPS). The associated skill score (CRPSS) is calculated relative to the CRPS for the ensemble-mean error climatology (ERR-CLI) forecast.
Figure 7: Probability integral transform (PIT) histograms (the analog of verification rank histograms for continuous forecast PDFs) as a function of nine ensemble variance categories (columns 1–9) for ERR-CLI, ENS-FIT, VAR-CEC, ENT-CEC, and MOD-CEC probabilistic error forecasts. PIT histograms for the total $10^4$-case sample are indicated in the far right column.
Figure 8: Probabilistic forecast sharpness using average 90% central prediction interval widths as a function of nine ensemble-variance categories for ERR-CLI, ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC error probability forecasts. The average central prediction interval widths are normalized by the climatological median ensemble standard deviation, $\sigma_F$. 
Figure 9: Predictive skill as a function of nine ensemble-variance categories for ENS-FIT, VAR-CEC, MOD-CEC, and ENT-CEC probability of success forecasts measured by the Brier skill score (BSS).