24.3 FORECASTING MESOSCALE UNCERTAINTY: SHORT-RANGE ENSEMBLE FORECAST ERROR PREDICTABILITY

Eric P. Grimit* and Clifford F. Mass
University of Washington, Seattle, Washington

1. INTRODUCTION

All types of end users who must make weather-dependent decisions stand to benefit greatly from knowing the expected accuracy of a particular forecast a priori. Forecast accuracy varies both spatially and temporally as a result of initial state and model errors, which change as the atmospheric flow evolves. Probabilistic weather forecasts derived from numerical weather prediction (NWP) ensembles can provide crucial information about the expected forecast uncertainty.

It has been theorized that ensemble spread should provide a measure of forecast uncertainty (Kalnay and Dalcher 1987; Murphy 1988; Houtekamer 1993), such that high (low) spread events correspond with high (low) forecast errors. The traditional approach to quantifying this so-called *spread-skill relationship* has involved finding the correlation between a measure of ensemble spread and the accuracy of a particular determinstic forecast. However, using the simple statistical arguments outlined by Houtekamer (1993, hereafter H93), there exists a theoretical limit to the strength of this type of correlation (Whitaker and Loughe 1998). From Eq. (33) in H93, the correlation between the standard deviation of perfect ensemble forecasts (σ) and the absolute error of the ensemble mean ($|\bar{E}|$) is limited by the temporal spread variability. As the temporal variability in spread (β) increases, so does the spread-error correlation. In the upper limit of spread variability,

$$\lim_{\beta\to\infty} \rho(\sigma, |\bar{E}|) = \sqrt{\frac{2}{\pi}}. \quad (1)$$

Therefore, spread-error correlation is not required to be large, even for an ensemble that properly samples all sources of uncertainty (Whitaker and Loughe 1998).

The theoretical limitations of spread-error correlation have been substantiated by empirical findings. Using various measures of spread and accuracy with both perfect (Barker 1991) and imperfect (Kalnay and Dalcher 1987; Molteni et al. 1996; Buizza 1997; Whitaker and Loughe 1998; Moore and Kleemann 1998; Hamill and Colucci 1998; Stensrud et al. 1999; Elsberry and Carr 2000; Goerss 2000; Ziehmann 2000; Hou et al. 2001; Grimit and Mass 2002; Stensrud and Yussouf 2003) models in a range of forecasting applications from tropical cyclone tracks to mesoscale convective precipitation, the relationship is usually highly scattered and correlation coefficients generally do not exceed 0.6–0.7. Potentially higher correlations can be achieved by considering only cases with extreme spread (Whitaker and Loughe 1998; Grimit and Mass 2002). Often the linear relationship between spread and accuracy is quite poor.

Some have concluded that variance-like measures of ensemble forecast spread are not the best predictors of forecast accuracy. For example, Toth et al. (2001) and Ziehmann (2001) suggest that categorical measures of forecast spread, such as mode population or statistical entropy are more skillful at discriminating between forecast successes and failures. A categorical analysis of this type requires that ensemble forecasts and corresponding verifications be divided into predetermined bins.

By expecting a linear relationship between ensemble spread and forecast accuracy, one assumes that forecast uncertainty and forecast error are equivalent. In fact, forecast uncertainty is better defined as the forecast error *distribution*, from which any observed forecast error can be considered a random variate. The trouble is that there is only one sample available, making it impossible to know the exact shape of that particular error distribution. One solution is to group forecast errors together from other cases and locations that potentially share some common characteristic. This grouping can be organized by ensemble forecast spread or another aspect of the ensemble forecast. The distribution of forecast errors in each group approximates the shape of the true average error distribution over all of those like samples.

When ensemble mean forecast errors are strati-

* *Corresponding author address:* Eric P. Grimit, Univ. of Washington, Dept. of Atmospheric Sciences, Seattle, WA 98195-1640; e-mail: epgrimit@atmos.washington.edu

fied according to their corresponding ensemble forecast variances, an error distribution associated with each variance bin is obtained. Essentially, ensemble mean forecast error can be viewed as a multi-valued function of ensemble forecast variance. Therefore, it is not surprising that ensemble forecast variances and ensemble mean forecast errors do not perfectly correspond, since the data pairs do not fit a straight line. On the other hand, it can be expected that ensemble forecast variance should perfectly correlate with the variance of the ensemble mean forecast error distributions (Wang and Bishop 2003).

Because of the need to consider the forecast error distribution, forecast error prediction should ideally be undertaken within a probabilistic framework. The traditional approach to forecast error prediction has been inherently deterministic, even though its basis is dervied from probabilistic considerations. When applying a least-squares regression, as in the spread-error correlation method, only the mean values of these error distributions are estimated. This deterministic limitation underscores the need for the problem to be defined in a fully probabilistic sense.

Skillful probabilistic error predictions can be derived directly from the ensemble, provided that a reasonably accurate forecast probability density function (PDF) can be produced. In practice, current ensemble forecasts are severely biased, uncalibrated, and limited in size, thus preventing accurate prediction of the true forecast PDF. These limitations are particularly pronounced for mesoscale prediction of surface weather variables (e.g. Eckel 2003). Unless these obstacles are overcome through superior ensemble forecast generation and statistical post-processing, an alternative method must be employed. Certainly, valuable uncertainty information is contained within imperfect ensemble predictions that could be utilized.

The goal of the present work is to more thoroughly investigate the ability of imperfect short-range ensemble forecasts (SREFs) to predict the mesoscale errors of sensible surface weather forecasts, such as those for wind and temperature. Previous work by the authors has involved the analysis of spread and error for near-surface wind direction forecasts from a University of Washington (UW) SREF system over the U.S. Pacific Northwest (Grimit 2001; Grimit and Mass 2002). The approach used to quantify forecast error predictability in those experiments has many limitations, including the fact that it is inherently deterministic. The aim here is to extend the analysis of mesoscale SREF spread-skill relationships to other measures of spread and accuracy, additional sensible surface weather parameters, and a probabilistic framework of forecast error prediction.

2. EXPERIMENTAL APPROACH

2.1 *A Simple Stochastic Model Framework*

A modified version of the H93 stochastic model was developed to establish the expected upper limit of both deterministic and probabilistic forecast error predictability. A large sample size of fictional ensemble predictions were generated stochastically, rather than dynamically as in real ensemble forecasts. The model was developed to account for the sampling effects of finite ensemble size and to allow for the use of varying measures of spread and accuracy. The simple model steps are as follows:

1. Draw the true forecast spread from a log-normal distribution as in the H93 model using,

   $$\ln(\sigma) \sim N(\ln(\bar{\sigma_f}), \beta^2), \qquad (2)$$

   with the mean forecast spread given by $\bar{\sigma_f}$ and β representing the standard deviation over time of the spread.

2. Explicitly simulate ensemble forecasts by drawing M values from the true distribution:

   $$F_i \sim N(Z, \sigma^2); \qquad i = 1, 2, \ldots, M, \qquad (3)$$

   where Z is the mean of the true distribution. Note that Z itself is drawn from a Gaussian climatological distribution with mean Z_c and variance σ_c^2, where $\sigma_c^2 \gg \bar{\sigma_f}^2$ (assuming short-range forecasts).

3. Draw the verification from the same distribution as the forecasts:

   $$V \sim N(Z, \sigma^2). \qquad (4)$$

4. Calculate the sample estimates of spread and accuracy. For example, using the variance (VAR) and absolute error of the ensemble mean (AEM),

   $$VAR = \frac{1}{M-1}\sum_{i=1}^{M}(F_i - \bar{F})^2; \qquad \bar{F} = \frac{1}{M}\sum_{i=1}^{M} F_i,$$

   $$AEM = |(\bar{F} - V)|.$$

5. Repeat the previous steps for many independent realizations.

The stochastic ensemble forecasts may be treated as perfect because the verification is drawn from the same distribution. The only built-in limitation comes from the finite sampling of the distribution, such that the potential for poor spread estimation is larger at smaller ensemble sizes. This mimics the situation found in real-world forecast error prediction, where the only estimate one has of the true spread is given by the variation among the ensemble forecasts at hand. Note that, the new model does not modify the H93 assumptions that the ensemble member forecasts are perfect and that the statistics are purely Gaussian.

The flexibility of this simple model allows for the calculation of categorical measures of spread and accuracy provided that the forecasts and verifications are partitioned into bins. The categories may be determined from climatology, have a fixed width, or correspond to critical thresholds that are user-dependent. Attention is restricted here to climatologically equally likely bins. Within this categorical framework, a measure of forecast uncertainty such as the mode population (M_{mode}) can be defined by

$$M_{mode} = \max(M_i); \qquad i = 1, 2, \ldots, nbin, \qquad (5)$$

where M_i is the number of ensemble forecasts contained in bin i and $nbin$ is the total number of bins. The modal frequency (MOD) is just M_{mode}/M, where M is the total number of ensemble members. Likewise, the statistical entropy (ENT) of a forecast distribution can be defined by

$$ENT = -\sum_{i=1}^{nbin} f_i \log_2 f_i, \qquad (6)$$

where f_i is the frequency of forecasts in bin i and ENT is measured in bits. Categorical forecast accuracy can be measured either with the ranked probability score (RPS) or with the Brier Score (BS) if no partial credit is to be given. In the latter case, a forecast is classified a success if the verification falls into the same bin as the ensemble mean (or mode) and a failure if it does not.

To approximate the idealized spread-skill relationship for perfect ensembles of finite size, 10000 fictional ensemble forecasts were simulated numerically using the simple model. Ensemble size (M) was varied from 2–50, which was sufficient to elucidate the asymptotic behavior of the spread-skill relationship with increasing ensemble size. The experiments were repeated for differing values of temporal spread variability (β) from 0.1–0.9 in 0.2 increments[1]. Both continuous (VAR) and categorical (ENT, MOD) measures of forecast spread were utilized. The association between spread and accuracy was measured by both the traditional spread-error correlation and the skill of probabilistic forecast error predictions using a cross-validation.

Probabilistic forecast accuracy was evaluated using the continuous ranked probability score (CRPS; Hersbach 2000) and the ignorance score (IGN; Roulston and Smith 2002). The baseline of comparison was the climatological error forecast, which was based on the full distribution of historical errors without any categorization. Forecast error predictability was then interpreted as the percentage improvement in CRPS/IGN over the baseline, which is identical to the associated skill scores (CRPSS/IGNSS).

2.2 *Probabilistic Forecast Error Prediction*

The skill of two methods of probabilistic forecast error prediction were evaluated. A method that does not require full knowledge of the true forecast PDF, which we called the conditional error climatology (CEC) method, uses the historical error distributions contained within bins organized by a common characteristic. The historical forecast errors were stratified according to their corresponding ensemble forecast uncertainty. In other words, a climatology of historical forecast errors, conditional on the ensemble forecast variance, was used as a probabilistic forecast error prediction. Forecast errors were also stratified by categorical measures of forecast uncertainty. The skill of the ensemble variance-based CEC method (VAR-CEC), modal-frequency-based CEC method (MOD-CEC), and statistical-entropy-based CEC method (ENT-CEC) were compared to the skill of the direct ensemble PDF (ENS-PDF) method within the perfect ensemble context of the simple model. To test the skill of the methods within a real-world context, similar comparisons were attempted using imperfect SREF data created at the University of Washington.

2.3 *Ensemble Forecast and Verification Data*

i The Two UW SREF Systems

An expansion of the original UW MM5 SREF system (Grimit and Mass 2002) took place in fall 2001 with the acquisition of additional large-scale analyses (Mass et al. 2003). As a result, two separate eight-member SREF systems were developed with detailed descriptions contained in Eckel (2003).

[1] Typically observed values of β from real dynamical forecast ensembles lie in the range 0.3–0.5

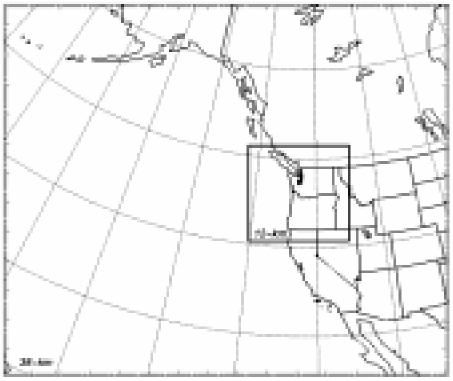

Figure 1. The 36- and 12-km MM5 domains for the University of Washington (UW) short-range ensemble forecast (SREF) systems.

The first system, called ACME^{core}, is a multianalysis, single-model (MM5) ensemble driven by initial conditions (ICs) and lateral boundary conditions (LBCs) obtained from major operational weather centers worldwide. The second system, called ACME^{core+}, is a multianalysis, perturbed-model ensemble driven by the same ICs/LBCs as in ACME^{core}, but also attempts to represent model uncertainty using the system simulation approach of Houtekamer et al. (1996). Both SREF systems were run at 36-km and 12-km horizontal resolution in a one-way nested configuration over the region shown in Figure 1. Only 12-km UW SREF output is analyzed here.

ii Post-Processing of UW SREF Data

Forecast biases comprised a significant component of the forecast error and impacted the estimation of forecast spread due to the disparate biases of the individual member forecasts. Because biases tend to be relatively consistent and predictable, it was logical to implement a bias-correction procedure that attempted to remove the majority of the bias before any other calculations were performed. A very simple, univariate bias-correction methodology was implemented (Eckel 2003). Forecast biases were calculated individually for each UW SREF member, each geographical location, and each forecast lead time. A moving-window training period was fixed to a length of 14 days, although various durations were tested. The optimal training period is likely state- and variable-dependent, however, that aspect has not been considered here.

The skill of the CEC probabilistic forecast error prediction method was compared to the direct PDF method both before and after bias-correction was applied. The direct PDF method requires that a smooth probability distribution be derived from the raw (or bias-corrected) forecast ensemble. This smoothing step can be considered a post-processing activity itself; one that is subject to a wide variety of methodologies. Ensemble forecasts can be smoothed by assuming a distribution and fitting the parameters. Of course, this fitted PDF is likely to be uncalibrated and underdispersive given the current limitations of ensemble forecasts. Alternative post-processing methods have been formulated that use historical ensemble member forecast errors to widen the forecast PDF and achieve calibration (e.g. Roulston and Smith 2002). A statistically principled way of combining forecasts from different sources and their historical errors to arrive at a calibrated and sharp forecast PDF is through Bayesian Model Averaging (BMA; Hoeting et al. 1999; Raftery et al. 200x). Probabilistic error forecasts were evaluated using both standard forecast PDF smoothing and BMA.

iii Verification Data and Evaluation Period

The NCEP 20-km Rapid Update Cycle (RUC20; Benjamin et al. 2003) analysis was used as truth in a grid-based verification approach. Observation-based verification has initially produced qualitatively similar results, but has suggested that a weaker spread-skill relationship existed. Evaluated data were taken from one cool season (October 2002 – March 2003). A total of 129 individual cases were selected during that period, each containing a complete set of SREF forecasts and verification data. For brevity, only results for grid-based verification of 2-m temperature forecasts are included in the present paper.

3. PRELIMINARY RESULTS AND DISCUSSION

3.1 *Simple Stochastic Model*

An intercomparison of the idealized predictive skill of four probabilistic forecast error prediction methods, given a perfect 50-member ensemble with temporal spread variability of 0.5, is presented in Figure 2. Three different measures of forecast spread were used to form the CEC predictions; one continuous (VAR) and two categorical (ENT and MOD). The predictive skill, measured by the continuous ranked probability skill score (CRPSS), is broken down by forecast spread. The largest predictive skill for all methods tends to be realized for cases with extreme (high or low) spread, reinforcing the findings of Houtekamer (1993), Whitaker and Loughe (1998),

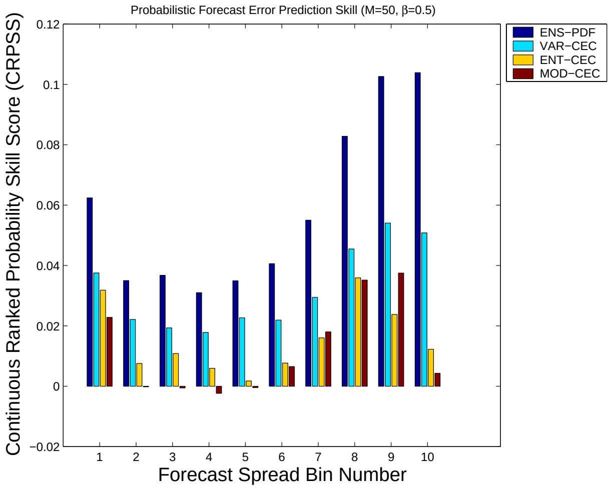

Figure 2. Idealized probabilistic forecast error predictive skill as a function of ensemble forecast spread for the direct ensemble PDF method (ENS-PDF) and the ensemble variance-based, modal frequency-based, and statistical entropy-based conditional error climatology methods (VAR-CEC, MOD-CEC, and ENT-CEC) measured by the continuous ranked probability skill score (CRPSS) and calculated from the simple stochastic model outlined in Section **2.1** with $M = 50$ and $\beta = 0.5$.

Table 1. CRPSS over the entire 10000-case sample from the idealized stochastic model for the ENS-PDF, VAR-CEC, MOD-CEC, and ENT-CEC methods of probabilistic forecast error prediction.

ENS-PDF	VAR-CEC	MOD-CEC	ENT-CEC
0.060	0.031	0.012	0.017

and Grimit and Mass (2002). Cases where the forecast spread is about average tend to display less forecast error predictability. In those average spread cases, the variances of the ensemble PDFs and the CEC distributions are likely to be more similar to the climatological error variance.

Given a perfect ensemble, ENS-PDF should be the most effective forecast error prediction method because of the perfect case-to-case resolution of the true forecast uncertainty. That hypothesis is verified by Figure 2, where ENS-PDF displays the largest idealized predictive skill for all forecast spread magnitudes. The aggregate CRPSS for the entire 10000-case sample (Table 1) is 0.06 for ENS-PDF and only 0.03 for VAR-CEC, the best-performing CEC approach as evaluated with this continuous measure. ENT-CEC performs slightly better than MOD-CEC, probably a reflection of the fact that ENT is calculated using the entire forecast distribution while MOD is not. Ignorance skill scores (IGNSS) indicate identical relative skill (not shown).

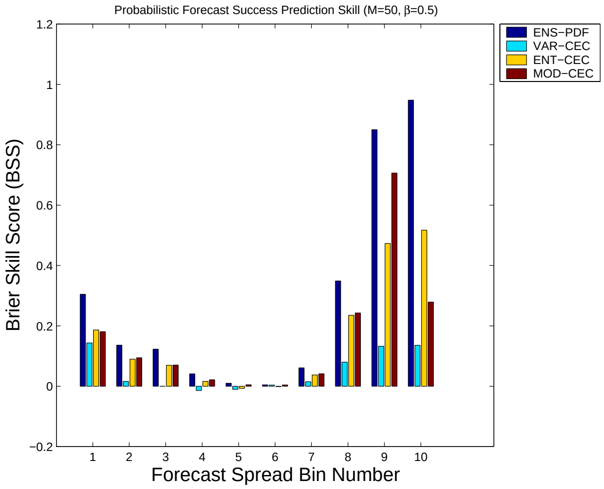

Figure 3. Idealized probabilistic forecast success predictive skill as a function of ensemble forecast spread for the ENS-PDF, VAR-CEC, MOD-CEC, and ENT-CEC methods as in Figure 2, except measured by the Brier skill score (BSS).

Table 2. BSS over the entire 10000-case sample from the idealized stochastic model for the ENS-PDF, VAR-CEC, MOD-CEC, and ENT-CEC methods of probabilistic forecast success prediction.

ENS-PDF	VAR-CEC	MOD-CEC	ENT-CEC
0.164	0.025	0.103	0.095

The relative perfomance of the CEC error prediction methods is somewhat different for end users who do not have a continuous utility function. After projecting the ensemble forecasts and verifications into ten climatologically equally likely bins, a probability of success was computed. In Figure 3 and Table 2, the Brier skill scores (BSSs) calculated relative to the climatological forecast success rate are reported. From this categorical vantage point, ENT-CEC and MOD-CEC outperform VAR-CEC. The ENS-PDF method remains the leader. Therefore, categorical (continuous) measures of forecast spread are more appropriate for end users with a categorical (continuous) sensitivity to forecast error. The drawback of using ENT and MOD as predictors of forecast error is that both measures are sensitive to ensemble size. At small ensemble sizes, there are only a small number of discrete values that ENT and MOD can take on, decreasing their ability to discriminate between events.

3.2 *UW SREF Systems*

To be presented at the 20thWAF/16thNWP conference.

3.3 *Effects of Post-Processing*

To be presented at the 20thWAF/16thNWP conference.

4. FUTURE WORK

Forecast error predictability using the direct PDF method after the application of BMA post-processing to the UW SREF forecasts is forthcoming. Additional seasons of UW SREF data will be evaluated, including one warm season (May – September 2003) and one cool season (October 2003 – March 2004). The forecast error predictability of the UW SREF and NCEP SREF systems will be compared over the October 2003 – March 2004 period.

Acknowledgements

The authors would like to thank F. Anthony (Tony) Eckel and M. Steven Tracton for their valuable comments and suggestions. Adrian Raftery and Tilmann Gneiting provided assistance with the implementation of Bayesian Model Averaging and the calculation of CRPS in the presence of a continuous PDF. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

References

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate*, **4**, 733–742.

Benjamin, S. G., D. Devenyi, S. S. Weygandt, K. J. Brundage, J. M. Brown, G. Grell, D. Kim, B. E. Schwartz, T. G. Smirnova, T. L. Smith, and G. S. Manikin, 2003: An hourly assimilation/forecast cycle: The ruc. *Mon. Wea. Rev.*, **131**, in press.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ecmwf ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119.

Eckel, F. A., 2003: *Effective short-range ensemble forecasting*. Ph.D. thesis, University of Washington.

Elsberry, R. L. and L. E. Carr, 2000: Consensus of dynamical tropical cyclone track forecasts—error versus spread. *Mon. Wea. Rev.*, **128**, 4131–4138.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193.

Grimit, E. P., 2001: *Implementation and evaluation of a mesoscale short-range ensemble forecasting system over the Pacific Northwest*. Master's thesis, University of Washington.

Grimit, E. P. and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the pacific northwest. *Wea. Forecasting*, **17**, 192–205.

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of eta-rsm ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, **14**, 382–401.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the samex '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.*, **121**, 1834–1846.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Kalnay, E. and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356.

Mass, C. F., M. Albright, D. Ovens, R. Steed, M. MacIver, E. Grimit, T. Eckel, B. Lamb, J. Vaughan, K. Westrick, P. Storck, B. Colman, C. Hill, N. Maykut, M. Gilroy, S. A. Ferguson, J. Yetter, J. M. Sierchio, C. Bowman, R. Stender, R. Wilson, and W. Brown, 2003: Regional environmental prediction over the pacific northwest. *Bull. Amer. Meteor. Soc.*, **84**, 1353–1366.

Molteni, F. R., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

Moore, A. and R. Kleemann, 1998: Skill assessment for enso using ensemble prediction. *Quart. J. Roy. Meteor. Soc.*, **124**, 557–584.

Murphy, J., 1988: Impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–493.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski: 200x, Calibrated mesoscale short-range ensemble forecasting using bayesian model averaging, unpublished manuscript, Department of Statistics, University of Washington.

Roulston, M. S. and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Stensrud, D. J., H. E. Brooks, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Stensrud, D. J. and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over new england. *Mon. Wea. Rev.*, **131**, 2510–2524.

Toth, Z., Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463–477.

Wang, X. and C. H. Bishop, 2003: A comparison of breeding and ensemble transform kalman filter ensemble forecast schemes. *J. Atmos. Sci*, **60**, 1471–1489.

Whitaker, J. S. and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302.

Ziehmann, C., 2000: Comparison of a single-model eps with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299.

— 2001: Skill prediction of local weather forecasts based on the ecmwf ensemble. *Nonlinear Processes Geophys.*, **8**, 419–428.