7.3B INVESTIGATION OF THE LINEAR VARIANCE CALIBRATION USING AN IDEALIZED STOCHASTIC ENSEMBLE

Walter C. Kolczynski, Jr.*
David R. Stauffer
Sue Ellen Haupt
Aijun Deng
Pennsylvania State University, University Park, PA

1. INTRODUCTION

The uncertainty in meteorological (MET) predictions is of great interest for a large number of applications, ranging from economic to recreational to public safety. It is therefore important that numerical models and forecasts provide accurate estimates of their uncertainty along with their best or most likely prediction (NRC 2007). One common method for determining this uncertainty is the use of ensembles, with multiple numerical forecasts produced using slightly different initial conditions and/or model parameterizations. The goal of using an ensemble is to span the possible outcomes given the uncertainties in the initial state of the atmosphere, the limited observations and the modeling system (Leith 1974). The mean of the ensemble has also been shown to often outperform any individual ensemble member compared to observations, even for low-level variables (e.g. Jones, Colle and Tongue 2007).

While ensemble forecasting is a significant step toward forecasting the most likely outcome and the uncertainty in the forecast, the size of operational ensembles is insufficient to fully represent the probability density function (PDF) of possible forecasts. An ensemble capable of doing so is impractical with current computing resources. Therefore, any MET ensemble provides a sampling of the full forecast PDF and any measures of the uncertainty from the ensemble (such as variance) should be evaluated for applicability and calibrated if necessary. Many studies, including Houtekamer et al. (1997), show that most MET ensembles are under-dispersive (the ensemble spread is consistently smaller than the spread in the forecast errors). Several studies attempt to determine a correlation between ensemble spread and some measure of the error for various variables (Kalnay and Dalcher 1987, Murphy 1988, Barker 1991, Houtekamer 1993, Buizza 1997, Hamill and Colucci 1998, Stensrud et al. 1999), but these studies generally find an unacceptably low correlation between spread and errors (Grimit and Mass 2007). Houtekamer (1993) explains this low correlation using a stochastic model that showed that, even in idealized cases, the correlation between ensemble spread and absolute error will not be large.

Grimit (2004) proposes that, rather than a simple spread-to-error correlation, a more probabilistic approach should be used to evaluate ensemble uncertainty. Specifically, Grimit stresses the distinction between forecast error and forecast uncertainty. Because each forecast results in only one realization of the forecast error (one random draw from the forecast PDF), the error of any particular forecast provides little information about the distribution from which it was drawn. However, if the relationship between the spread of the MET ensemble and forecast uncertainty is constant within a sample, we can group multiple samples with similar ensemble variance and use the distribution of the associated group of errors as a measure of the underlying uncertainty in the errors.

Grimit (2004) and Grimit and Mass (2007) use an idealized stochastic model that shows a strong linear relationship between ensemble spread and error variance. Kolczynski et al. (2009) uses a similar method derived from Roulston (2005) on low-level wind data from the National Centers for Environmental Prediction Short-Range Ensemble Forecast (NCEP-SREF). Kolczynski et al. (2009) also finds a strong correlation between ensemble variance and error variance, and uses this linear relationship as a calibration (denoted Linear Variance Calibration or LVC) for wind variance input into an atmospheric transport and dispersion model. However, unlike the Grimit and Mass study using an idealized model, Kolczynski et al. (2009) finds that the slope of the linear fit is less than the ideal value of one, and that the y-intercept is larger than the ideal value of zero (Fig. 1). The study also finds that these parameters change substantially depending on the length of the forecast. When applied to dispersion calculations in a case study, the calibration improves the resulting dispersion forecast in most of the metrics used in the study.

This study further explores possible influences on the LVC slope and intercept. This is done using a stochastic model adapted from that used by Houtekamer (1993) and Grimit and Mass (2007). However, this new stochastic model allows the variance of the error distribution and the ensemble distribution to vary from each other, allowing for the creation of “imperfect” ensembles, but in such a manner that a linear relationship between the variances is maintained. Since we are interested in the variances of low-level wind speed for atmospheric transport and dispersion, the new model also uses a Weibull distribution instead of a log-normal distribution as the underlying “climatology”, as a Weibull approximates the climatology of surface (10-m) wind speed (Wilks 2006). Section 2 provides the details of the new stochastic model. Section 3 presents the results of the model for six

* *Corresponding Author Address:* Walter Kolczynski, Jr. Dept. of Meteorology, Pennsylvania State University, 503 Walker Bldg. State College, PA 16802; wck122@psu.edu.

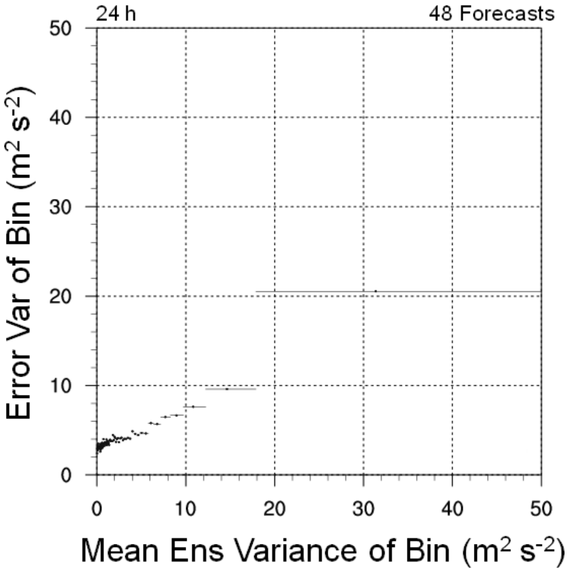

Figure 1: Scatterplot showing the relationship between error variance and ensemble variance (adapted from Kolczynski et al. 2009). The linear best-fit line from the LVC (not shown) has a slope of 0.502 and a y-intercept of 3.007.

experiments using an ensemble size of 20, typical of operational MET ensembles. The same six experiments are then repeated using varying ensemble size for results in Section 4. Section 5 details a theoretical approach to determining the variations in LVC slope with ensemble size. Conclusions and directions for future work are offered in Section 6.

2. METHODOLOGY

In order to investigate the relationship between ensemble variance and error variance, we construct a stochastic model in which we control the variance from which the data are drawn. This will allow us to compare the results of the model directly with the expected values given the underlying distribution.

For each simulation, we first generate a random value of "speed" s_i from a Weibull distribution with a shape parameter of 1.8 and a scale parameter of 5.0. We choose this distribution because we are interested in the low-level winds important for atmospheric transport and dispersion, and Kolczynski et al. (2009) focused on low-level winds. The shape and scale parameters fall within the common range of values empirically determined for the distribution of wind speeds. The error distribution and ensemble distribution are then both defined to be normal distributions with a mean of zero and a variance that depends on s_i. The relationship is a simple linear one, with the variance of the errors (σ_a^2) defined as

$$\sigma_a^2 = m_a \cdot s_i + b_a \tag{1}$$

and the variance of the ensemble (σ_e^2) defined as

$$\sigma_e^2 = m_e \cdot s_i + b_e \,. \tag{2}$$

Because the variance of the product of normals is the sum of the individual normal variances, the b_a in (1) can also be thought of as the observation error. Using these variances, we pick one number at random from the normal distribution $N(0,\sigma_a^2)$ as our observed error, and M numbers from the distribution $N(0,\sigma_e^2)$ to serve as our ensemble. This process is repeated one-hundred thousand times to create a population of error/ensemble pairs.

A modified version of the LVC presented in Kolczynski et al. (2009) is then used to calculate a relationship between the ensemble variance and error variance. First, the error-ensemble pairs are ordered based on the ensemble variance. Then the data are binned into groups of 1000. In each bin, the ensemble variance is averaged to obtain a representative ensemble variance for the bin, and the variance of the errors in the bin is computed. The philosophy is that, if the relationship between ensemble variance and error variance remains constant within the sample, errors from cases with similar ensemble variance should be drawn from similar error distributions. Thus, we can take the variance of the errors from many different realizations and it would be similar to the variance if we could compute it over many errors for the same case. We then perform a linear regression on the mean ensemble variance/error variance pairs for each bin to determine the slope ($\hat{m}_{LVC}$) and intercept of the relationship ($\hat{b}_{LVC}$).

Because we specify the underlying distribution, we can also compute the expected LVC slope and intercept algebraically. The slope of the LVC regression (m_{LVC}) is expected to be

$$m_{LVC} = \frac{m_a}{m_e} \tag{3}$$

and the y-intercept of the LVC regression (b_{LVC}) is expected to be

$$b_{LVC} = b_a - b_e \frac{m_a}{m_e} \,. \tag{4}$$

We can compare these theoretical values to those calculated from the simulated data.

3. EXPLORATION WITH ENSEMBLE SIZE OF 20

In order to explore the behavior of the model with typical operational ensemble sizes, we first consider several configurations using a constant ensemble size of 20 members. Each experiment is summarized in Table 1 along with the 20-member ensemble results. Experiments A and B are both "perfect" ensembles, with the ensemble being drawn from a distribution identical to that from which the errors are drawn. The only difference between A and B is that the variances of B are three times bigger than those of A relative to s_i. Experiments C and D explore simple over- and under-dispersive cases, with the ensemble being drawn from distributions with three times larger or one third smaller variances, respectively. Experiments E and F

Table 1: Summary of the experimental configuration and corresponding results for an ensemble size of 20. The color of each row corresponds to the color used in Figs. 3 and 4.

Exp.	m_a	b_a	m_e	b_e	R^2	m_{LVC}	$\widehat{m}_{LVC}$	b_{LVC}	$\widehat{b}_{LVC}$
A	0.1	0.0	0.1	0.0	0.94	1.00	0.71	+0.00	+0.13
B	0.3	0.0	0.3	0.0	0.95	1.00	0.72	+0.00	+0.38
C	0.1	0.0	0.3	0.0	0.98	0.33	0.24	+0.00	+0.13
D	0.3	0.0	0.1	0.0	0.96	3.00	2.15	+0.00	+0.39
E	0.1	0.0	0.1	0.4	0.99	1.00	0.45	- 0.40	+0.07
F	0.1	0.4	0.1	0.0	0.92	1.00	0.72	+0.40	+0.52

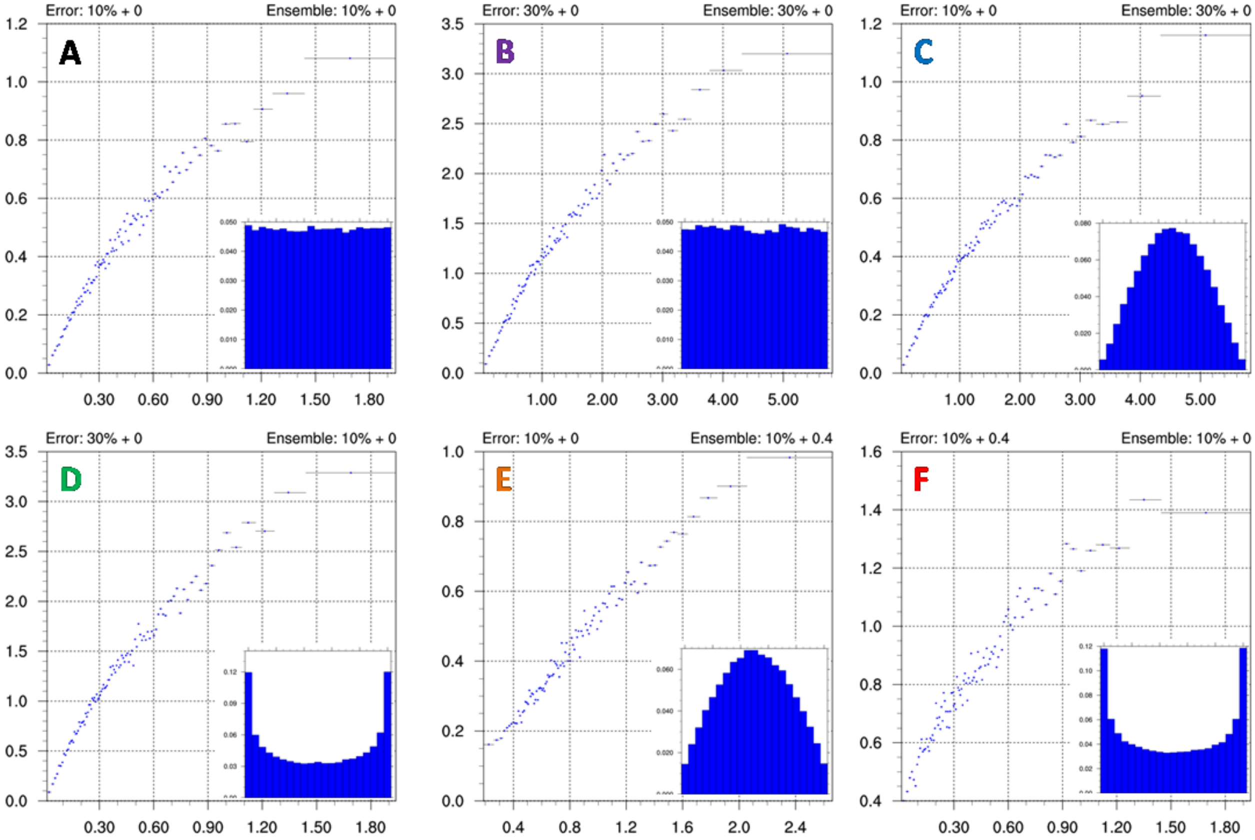

Figure 2: Relationship of ensemble variance (abscissa) to the error variance (ordinate) for each experiment, using an ensemble size of 20. Insets: rank histogram of each experiment

demonstrate more complicated relationships, where the variance of the error distribution or ensemble distribution includes a constant. Figure 2 shows the scatterplot and rank histogram from each experiment.

The most important result is that the slope calculated using the LVC ($\widehat{m}_{LVC}$) is smaller than the slope expected from algebra (m_{LVC}). Similarly, the y-intercept computed using LVC ($\widehat{b}_{LVC}$) is higher than the theoretical value from algebra (b_{LVC}). However, the R^2 for every experiment is high (above 0.92), indicating a strong linear correlation between ensemble variance and error variance even though the coefficients of the linear fit do not match the theoretical values. Interestingly, the slope computed by LVC is 72% of the theoretical value for every experiment except E, which is the only one to use an additive constant for the ensemble variance. There also seems to be a functional relationship for the y-intercept calculated by LVC, with the LVC y-intercept being $1.3 \cdot m_a$ larger than the theoretical value for every experiment except E.

4. EXPLORATION OF ENSEMBLE SIZE

To investigate the possible effect of ensemble size on the ensemble variance/error variance relationship, particularly the deviation from theoretical values, we repeat each experiment for a variety of ensemble sizes ranging from five to five thousand (5, 10, 20, 50, 100, 200, 300, 400, 500, 1000 and 5000). This should reveal any sampling limitations due to ensemble size, as well as indicate any fundamental problems with the LVC if it is unable to obtain the theoretical values at very large ensemble sizes.

Figures 3 and 4 show the calculated LVC slope and y-intercept respectively for each experiment at varying ensemble sizes. The LVC-calculated values of slope and y-intercept for each experiment approach the theoretical value as ensemble size gets larger. This indicates that the deviation from the theoretical values in the experiments when using 20 members is likely due to sample size issues and is not a byproduct of LVC. Furthermore, the plot shows that much larger ensemble sizes than those currently used operationally are needed (> 200) in order to obtain the theoretical slope and y-intercept. This is an important outcome, because it means that *even with a perfect ensemble*, the ensemble variance should be calibrated for any ensemble with fewer than several hundred members. Additional "climatological" distributions of the wind speed were also considered (not shown) with similar results. Scientists exploring applications of ensemble variance ranging from ensemble generation to its use as a proxy for uncertainty should be mindful of this result.

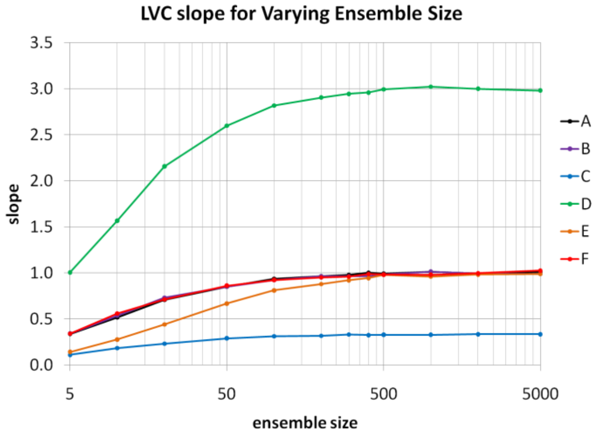

Figure 3: Calculated LVC slope for each of the experiments listed in Table 1 using variable ensemble sizes. The theoretical slope is 1 for Exps. A, B, E and F; $\frac{1}{3}$ for Exp. C; and 3 for Exp. D. Data are plotted at ensemble sizes of 5, 10, 20, 50, 100, 200, 300, 400, 500, 1000, 2000 and 5000.

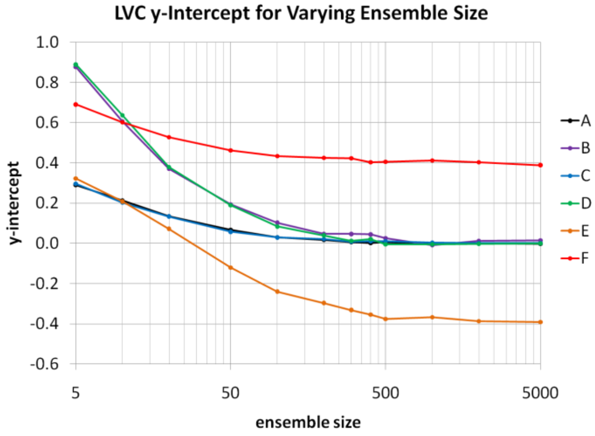

Figure 4: Calculated LVC y-intercept for each of the experiments listed in Table 1 using variable ensemble sizes. The theoretical y-intercept is 0 for Exps. A-D; -0.4 for Exp. C; and 0.4 for Exp. D. Data are plotted at ensemble sizes of 5, 10, 20, 50, 100, 200, 300, 400, 500, 1000, 2000 and 5000.

5. RESEMBLANCE TO ERROR-IN-VARIABLES MODEL

The changing behavior of the LVC-estimated slope and intercept bring to mind the error-in-variables model (Casella and Berger, 2001). In the error-in-variables model, two random variables X_i and Y_i have expected values that have a linear relationship. The expected values of variables X_i and Y_i are often called latent variables and can be thought of as the "true" value subject to measurement error. The system of equations becomes

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \varepsilon_i \\ X_i &= \xi_i + \delta_i \end{aligned} \qquad (5)$$

where α and β are parameters of the linear relationship, ξ_i is the expected value of X_i, $\alpha + \beta\xi_i$ is the expected value of Y_i, and ε_i and δ_i are random variables with an expected value of zero (the "measurement error"). Commonly, ε_i and δ_i are considered to be normally distributed.

If a linear regression is performed for Y on X, the resulting estimates of the parameters will be biased, with the estimated slope, $\hat{\beta}$, given by

$$\hat{\beta} = \beta \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2}\,. \qquad (6)$$

Since variance is always positive, this means that the estimated slope is biased lower than the actual slope: this is commonly referred to as attenuation. If we assume that only σ_ξ^2 varies with ensemble size M and that it varies as the inverse of ensemble size so that it approaches zero as the ensemble size becomes infinite, we can condense all of the non-parameter constants to a single constant γ:

$$\hat{\beta} = \beta \frac{1}{1 + \dfrac{\gamma}{M-1}}\,. \qquad (7)$$

We can rearrange (7) into a linear relationship and substitute our previous notation for the beta parameters:

$$\frac{\beta}{\hat{\beta}} - 1 = \frac{m_{LVC}}{\hat{m}_{LVC}} - 1 = \frac{\gamma}{M-1} \qquad (8)$$

where we can estimate γ by regression. The ratio $\hat{\beta}/\beta$ is often called the *attenuation factor*. Thus this regression relates the attenuation factor to the ensemble size.

For our situation however, the random variable X does not have normal (Gaussian) errors. Instead, if we ignore the effect of the binning, the sample variance calculated from a random sample of points taken from a normal distribution (the ensemble) is actually a random variable proportional to a chi-squared distribution with degrees of freedom equal to the sample size minus one:

$$S_i^2 = \frac{\sigma_i^2}{n-1}\chi_{n-1}^2 \tag{9}$$

where S^2 is the sample variance, σ^2 is the true variance, and n is the sample size. Eq. (9) has an expected value of σ^2 and a variance of $2\sigma^2$.

Despite the different form of the random variable X, the regression of Y on X for our case can still be reduced to the form of (8), (but with a different constant γ). For each experiment except E, it will be the same constant, and using the results from experiments A-D and F in a single regression yields an estimate for γ of 7.865. If we insert this estimate back into (8) to "correct" the estimated LVC slope to the true value for these experiments, we find that the corrected estimates are very near the true value for all ensemble sizes (Fig. 5).

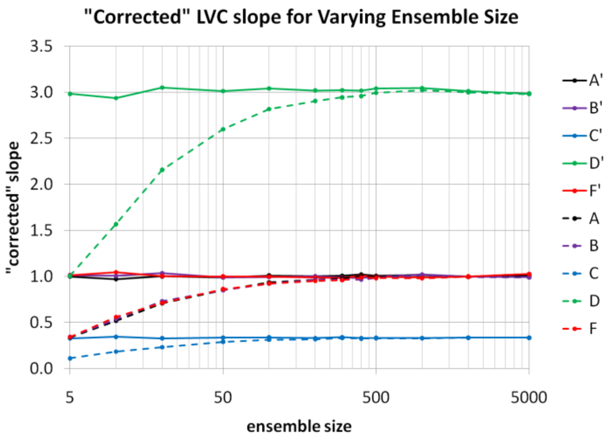

Figure 5: Dashed lines are as in Fig. 3 (excluding experiment E). Solid lines represent LVC slope values corrected for attenuation by sampling error in the ensemble variance for the same experiments.

6. CONCLUSIONS

This study has explored the relationship between ensemble variance and error variance calculated by the Linear Variance Calibration (LVC) presented in Kolczynski et al. (2009) in a controlled way using an idealized stochastic ensemble. The results show that the calculated LVC slope and y-intercept deviate substantially from the algebraically-derived values when ensemble size is less than several hundred members. This result implies that ensemble variances, even from otherwise "perfect" ensembles, should be calibrated if ensemble size is less than several hundred members.

This study also derived a linear relationship between the inverse of ensemble size and the inverse of the LVC slope attenuation due to ensemble size. This relationship is valid whenever b_e, the portion of the ensemble variance independent of the forecast value s_i, is zero. Using linear regression, we estimated the constant for the linear relationship, then used the relationship as a correction to the LVC-determined slope to the true slope. Corrected slopes fell close to the true LVC slopes at all ensemble sizes in each experiment where b_e is equal to zero.

Further research will determine if this relationship can be generalized for cases where b_e is not equal to zero. We will also determine if this method can be applied to cases where we don't know the true LVC slope. This would allow us to then apply the correction method to real-data LCV slopes. Such a correction would more accurately determine whether an ensemble has an appropriate error/spread relationship and would allow for a fair comparison between ensembles of different sizes.

Acknowledgements. We would like to thank our Penn State colleagues Naomi Altman, Arthur Small, and George Young for insightful conversations about uncertainty, probability and statistics that were instrumental in evaluating the stochastic model. This work was supported by the Defense Threat Reduction Agency (DTRA) under the supervision of Dr. John Hannan via Global Security & Engineering Solutions, a Division of L-3 Services, Inc., under subcontract 2008-1113, Prime Contract # DTRA01-03-D-0013.

7. REFERENCES

Barker, T.W., 1991: The Relationship Between Spread and Forecast Error in Extended-range Forecasts. *Journal of Climate*, **4**, 733-742.

Buizza, R., 1997: Potential Forecast Skill of Ensemble Prediction and Spread and Skill Distributions of the ECMWF Ensemble Prediction System. *Monthly Weather Review*, **125**, 99-119.

Casella, G. and R.L. Berger, 2001: *Statistical Inference, Second Ed.* Duxbery, Pacific Grove, CA, 660 pp.

Grimit, E.P., 2004: Probabilistic Mesoscale Forecast Error Prediction Using Short-Range Ensembles. Ph.D. dissertation, University of Washington, 146 pp.

------ and C.F. Mass, 2007: Measuring the Ensemble Spread-Error Relationship with a Probabilistic Approach: Stochastic Ensemble Results. *Monthly Weather Review*, **135**, 203-221.

Hamill, T.M. and S.J. Colucci, 1998: Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts. *Monthly Weather Review*, **126**, 711-724.

Houtekamer, P.L., 1993: Global and Local Skill Forecasts. *Monthly Weather Review*, **121**, 1834-1846.

------, J. Derome, H. Ritchie, and H.L. Mitchell, 1997: A System Simulation Approach to Ensemble Prediction. *Monthly Weather Review*, **125**, 3297-3319.

Jones, M.S., B.A. Colle and J.S. Tongue, 2007: Evaluation of a Mesoscale Short-Range Ensemble Forecast System over the Northeast United States. *Weather and Forecasting*, **22**, 36-55.

Kalnay, E. and A. Dalcher, 1987: Forecasting Forecast Skill. *Monthly Weather Review*, **115**, 349-356.

Kolczynski, Walter C. Jr., D.R. Stauffer, S.E. Haupt and A. Deng, 2009: Ensemble Variance Calibration for Representing Meteorological Uncertainty for Atmospheric Transport and Dispersion Modeling. *Journal of Applied Meteorology and Climatology*, **48**, 2001-2021.

Leith, C.E., 1974: Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, **102**, 409-418.

Murphy, J.M., 1988: The Impact of Ensemble Forecasts on Predictability. *Quarterly Journal of the Royal Meteorological Society*, **114**, 463-493.

National Research Council, 2007: Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. The National Academies Press, Washington, DC, 124 pp.

Roulston, M.S., 2005: A Comparison of Predictors of the Error of Weather Forecasts. *Nonlinear Processes in Geophysics*, **12**, 1021-1032.

Stensrud, D.J., H.E. Brooks, J. Du, M.S. Tracton and E. Rogers, 1999: Using Ensembles for Short-Range Forecasting. *Monthly Weather Review*, **127**, 433-446.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences, Second Ed.* Academic Press, Burlington, MA, 627 pp.