5.2 TAKING INTO ACCOUNT THE RANK OF A MEMBER WITHIN THE ENSEMBLE FOR PROBABILISTIC FORECASTING BASED ON THE BEST MEMBER METHOD

Vincent Fortin*
Meteorological Service of Canada, Dorval, Québec

Anne-Catherine Favre
INRS-ETE, Québec City, Québec

1. INTRODUCTION

Optimal use of an atmospheric forecast typically requires some information on its sharpness and reliability. Outputs from an ensemble prediction system (EPS) can be used to estimate both characteristics (Sivillo et al., 1997), but on the other hand do not prove to be very useful probabilistic forecasts, in part because the number of ensemble members is typically fairly small, but mainly because the forecasts are not perfectly reliable: they can be biased and typically do not display enough variability, thus leading to an underestimation of the uncertainty (Buizza et al., 2005).

Different approaches have been proposed recently to build reliable probabilistic forecasts from such ensembles, including Bayesian model averaging (Raftery et al., 2005), the Bayesian processor of output (Krzysztofowicz, 2004) and the best member method (Roulston and Smith, 2003), which is by far the simplest to implement. While we agree with Krzysztofowicz (2004) that Bayesian theory provides the appropriate theoretical framework for obtaining the probability distribution of a predictand, conditional on an ensemble of model outputs, we do feel that there is at the moment a need for simpler methods which can be readily implemented.

The best member method proposed by Roulston and Smith (2003) relies on a simple resampling scheme: individual members of an ensemble are "dressed" with an error distribution derived from a database of past errors made by the "best" member of the ensemble, where the best member of an ensemble forecast is defined as the ensemble member which described best the observed state of the atmosphere for this forecast, with respect to a given norm.

Wang and Bishop (2005) have however shown by stochastic simulations that the best member method can lead both to underdispersive and overdispersive ensembles. They proposed an improved method where the error distribution is scaled so as to obtain ensembles which display the desired variance. This approach however fails in cases where the undressed ensemble members are already overdispersive.

We propose to overcome this difficulty by dressing and weighing each member differently. This method leads to forecasts which not only have the right amount of variance both for underdispersive and overdispersive EPS, but which also have better tail behaviour. In this extended abstract, we only present a summary of the methodology and of the results. A more in-depth analysis of this idea can be found in Fortin and Favre (2005).

2. EPS WITH EXCHANGEABLE MEMBERS

A numerical weather prediction system provides forecasts using some information on the state of the atmosphere and a dynamical model of the atmosphere. To assess the impact of uncertainty on the analysis and on the dynamical model structure, it is possible to run different versions of a model (or different models altogether) with slightly different initial conditions, and thus obtain an ensemble of forecasts. We refer to such a system as an ensemble prediction system, or EPS.

In some cases, members of an EPS are (finitely) exchangeable, meaning essentially that there is no possibility that one can tell from the forecasts which version of the model was used to obtain a given forecast, and which method was used to perturbate the initial conditions. In other words, the forecasts can be reordered and renumbered without loosing any information on the predictand. For a formal definition of

**Corresponding author address :* Vincent Fortin, Numerical Prediction Research Division, Meteorological Service of Canada, 2121 Trans-Canada Highway, 5th Floor, Dorval, Québec, Canada H9P 1J3; e-mail : vincent.fortin@ec.gc.ca

exchangeability the reader may refer to the monograph by Bernardo and Smith (1994).

In many cases, even if the members of an EPS are not strictly exchangeable, they can be considered exchangeable in practice. For example, if a control run using unperturbed initial condition is included in the ensemble, it might be possible to distinguish it from the other members of the ensemble for the first few days of the forecast, so the members would not be exchangeable for short lead times, but as the lead time increases this typically becomes almost impossible, making the forecasts exchangeable in practice.

In the remaining of this paper, we will restrict ourselves to EPS for which the members can be considered exchangeable.

3. UNIVARIATE FORECASTING

In this paper, we focus on the problem of univariate forecasting from an EPS. Clearly, in most cases the predictand is multidimensional. In many cases however, this is because the variable of interest is really a combination of different meteorological elements at different spatial locations and lead times. This is the case for example for hydrological forecasting, where the inputs are certainly multidimensional, but the output is very often unidimensional. The proposed methodology would also be applicable in these circumstances. An extension of the methodology to multivariate forecasting is discussed in the last section of the paper.

4. PROBABILISTIC FORECASTING

Let y_t be the observation of interest at time t. For simplicity, we assume that $\{y_t\}$ is a realization of a stationary stochastic process and that forecasts are issued for a single lead time, so that they can be indexed simply by the time for which they are valid. Let $x_{t,k}$ be a forecast of y_t provided by the kth member of an EPS and $\mathbf{x}_t=\{x_{t,k}\}$ denote the ensemble of all forecasts of y_t.

From the EPS outputs $\mathbf{x}_t$, we would like to be able to obtain a probabilistic forecast $p(y_t|\mathbf{x}_t)$ of y_t, or at least to be able to simulate from this predictive distribution.

5. THE BEST MEMBER METHOD

Roulston and Smith (2003) have recently introduced the “best member” method to obtain scenarios which appear to be sampled from the predictive distribution $p(y_t|\mathbf{x}_t)$. The idea is to “dress” each ensemble member $x_{t,k}$ with a probability distribution representing the error made by this member when it happens to be the best member of the ensemble. Amongst the ensemble members $x_{t,k}$, define the best member x_t^* as the one which minimizes $\left|y_t - x_{t,k}\right|$.

From a database of past forecasts, build a probability distribution from realizations of $\varepsilon_t^*=y_t-x_t^*$. Finally, define a probabilistic forecast which consists of a finite mixture of error distributions centered on each original member of the ensemble:

$$\hat{p}(y_t \mid \mathbf{x}_t) = \frac{1}{K}\sum_{k=1}^{K} p_{\varepsilon^*}(y_t - x_{t,k}) \qquad (1)$$

Simulating from this distribution is simple: to obtain $M=N{\cdot}K$ simulations, dress each of the K members of the ensemble by sampling N error patterns $\{\varepsilon_{t,k,n}, n=1,2,...,N\}$ from a database of past forecast, and add each error pattern to this ensemble member:

$$\hat{y}_{t,k,n} = x_{t,k} + \varepsilon_{t,k,n} \qquad (2)$$

We refer to the initial ensemble $\mathbf{x}_t=\{x_{t,k}\}$ as a dynamical ensemble, and to the augmented ensemble $\hat{\mathbf{y}}_t=\{\hat{y}_{t,k,n}\}$ as a statistical ensemble.

Wang and Bishop (2005) have shown that the best member method does not lead to reliable forecasts. A system is said to be reliable if we cannot tell apart a large set of observed values from a large set of forecasts.

6. RESCALING THE ERROR PATTERNS

Wang and Bishop (2005) suggest that the error patterns be rescaled by a constant factor ω so that the variance of the statistical ensemble be the same as the variance of the observations.

$$\hat{y}_{t,k,n} = x_{t,k} + \omega \cdot \varepsilon_{t,k,n} \qquad (3)$$

This technique is only applicable when the dynamical ensemble is underdispersive. Indeed, as we still add uncorrelated noise to the dynamical ensemble, we can only increase the variance. In the conclusion of their paper, Wang and Bishop (2005) mention that a possible solution would be to weigh each ensemble member differently. We think that ensemble members should not only be weighed, but also be dressed differently.

7. IMPROVING UPON THE BEST MEMBER METHOD BY DRESSING AND WEIGHING EACH MEMBER DIFFERENTLY

In the best member method, each ensemble member is dressed using the same error distribution. This seems to make sense if all ensemble members are exchangeable prior to their observation, as there is no *a priori* reason for assuming that any ensemble member has any more chance of being the best member of the ensemble, or that its error distribution should be any different from that of the other members if it indeed happens to be the best member. This intuition is simply wrong.

7.1 *Why weigh or dress differently exchangeable members of an EPS?*

Let $\pi_k = \Pr(x_{t,k} = x_t^* \mid \mathbf{x}_t)$, i.e. the posterior probability that member k be the best member of the ensemble for time t. From the law of total probability, we have:

$$p(y_t \mid \mathbf{x}_t) = \sum_{k=1}^{K} \pi_k \cdot p(y_t \mid x_{t,k} = x_t^*, \mathbf{x}_t) \qquad (4)$$

Hence, the best member method would be exact, i.e. $\hat{p}(y_t \mid \mathbf{x}_t)$ would be equal to $p(y_t \mid \mathbf{x}_t)$ if $\pi_k = 1/K$ and $p(y_t \mid x_{t,k} = x_t^*, \mathbf{x}_t) = p_{\varepsilon^*}(y_t - x_{t,k})$ for k=1,...,K.

Exchangeable members of an EPS, being indistinguishable, have the same prior probability of being the best member of the ensemble, i.e. $\Pr(x_{t,k}=x_t^*)=1/K$. But this does not mean that the posterior probabilities $\Pr(x_{t,k} = x_t^* \mid \mathbf{x}_t)$ should be equal. Consider for example an EPS which produces underdispersive dynamical ensembles. Because they are underdispersive, the outcome will often lie outside the spread of the ensemble. Hence, the extrema of the ensemble will have much more chance of being the best member of the ensemble than a member which is close to the ensemble mean. Conversely, if the EPS is highly overdispersive, members close to the ensemble mean will have more chance of being the best members of the ensemble than the extrema of the ensemble.

The distribution $p(y_t \mid x_{t,k} = x_t^*, \mathbf{x}_t)$ also depends on k. Let $x_{t,(k)}$ be the k^{th} smallest forecast of the ensemble (or k^{th} order statistic) and let $m_{t,(k)}=(x_{t,(k\text{-}1)}+x_{t,(k)})/2$ be the mid-point between $x_{t,(k\text{-}1)}$ and $x_{t,(k)}$. The event $x_{t,(k)}=x_t^*$ occurs if and only if y_t belongs to the closed interval $[m_{t,(k)},m_{t,(k+1)}]$ for k=2,...,K-1. For k=1, $x_{t,(k)}=x_t^*$ occurs if y_t belongs to the half-opened interval $(-\infty,m_{t,(2)}]$, and for k=K, $x_{t,(k)}=x_t^*$ occurs if y_t belongs to the half-opened interval $[m_{t,(K)},+\infty)$. Hence, the support of $p(y_t \mid x_{t,k} = x_t^*, \mathbf{x}_t)$ is a bounded interval for all members except for the smallest and largest, for which it is semi-infinite. It is therefore impossible to find a distribution p_{ε^*} such that $p(y_t \mid x_{t,k} = x_t^*, \mathbf{x}_t) = p_{\varepsilon^*}(y_t - x_{t,k})$ for all k, since the support of p_{ε^*} does not depend on k.

The above examples show that the probability that an ensemble member be the best member of the ensemble as well as the error distribution of the best member of the ensemble both depend on the location of the ensemble member within the ensemble.

7.2 *Dressing and weighing each order statistic of the dynamical ensemble differently*

Given the obvious dependence on the rank of a member in an ensemble of the probability that it be the best member and of its error distribution if it is the best member, we propose to dress each order statistic of the dynamical ensemble differently, and to give them different weights when constructing a statistical ensemble.

Define $\boldsymbol{\varepsilon}_{(k)}{}^* = \{y_t - x_t{}^* \mid x_{t,(k)} = x_t{}^*, t = 1,2,\ldots,T\}$ to be the best member errors observed in the database of past forecasts when the best member was the k^{th} order statistic. Define also $\pi_{(k)}$ to be the probability that the best member be $x_{t,(k)}$, i.e. $\pi_{(k)} = \Pr(x_{t,(k)} = x_t^* \mid \underline{x}_t)$.

When dressing the kth member in the ordered ensemble, we propose that instead of resampling from the archive of all best member errors, we instead resample from $\boldsymbol{\varepsilon}_{(k)}{}^*$ to obtain dressed ensemble members.

The number of dressed members generated from the kth order statistic should reflect the probability that this particular member be the best member of the ensemble. If we want to obtain M ensemble members from the original K members, then a possibility is to draw $N_k=\pi_{(k)}\cdot M$ dressed ensemble members from $x_{t,(k)}$. However, as we are still not really drawing the statistical members from the conditional distribution $p(y_t|\underline{x}_t)$, we would have little chance to get statistical members which have exactly the right amount of variance.

An alternative is to optimize the number of dressed ensemble members drawn from each dynamical member so as to get the correct variance. Of course, many combinations of weights can lead to the same variance, so we have to constrain the problem. A solution is to choose a parametric function for w_k having a single parameter, which can be tuned to ensure that the forecasts have the same variance as the observations, for example a symmetric beta probability distribution function:

$$w_k \propto \int_{(k-1)/K}^{k/K} u^{\alpha-1} \cdot (1-u)^{\alpha-1}\, du \qquad (5)$$

8. EVALUATING THE RELIABILITY OF THE BEST MEMBER METHOD USING A SYNTHETIC EPS SYSTEM

To assess the reliability of the best member method, we shall re-use in this paper the simple synthetic EPS setup proposed by Wang and Bishop (2005).

8.1 *A synthetic EPS system*

Assume that observations $\{y_t,\ t=1,2,...,T\}$ are independent, normally distributed random variables with zero mean but time-dependent variance σ_t^2. Assume also that σ_t^2 is drawn from a Chi-square distribution with 3 degrees of freedom. An EPS provides a K-member forecast $\mathbf{x}_t=\{x_{t,k},\ k=1,2,...,K\}$ for each observation y_t. All ensemble members are independent, identically distributed (i.i.d.) normal variates having zero mean and time-dependent variance ξ^2_t, where ξ^2_t is related to σ_t^2 by a random relationship: $\xi^2_t=a_t\cdot\sigma_t^2$, a_t being a uniform random variable on the interval $[\mu_a\text{-}0.5,\mu_a+0.5]$, where μ_a is the expectation of a_t.

This EPS has spread-skill: when the observation y_t is less variable thus more predictable, i.e. σ_t^2 happens to be small, the variance of the ensemble members ξ^2_t will tend to be smaller, and conversely. The EPS will be underdispersive if $\mu_a<1$ and overdispersive if $\mu_a>1$.

To evaluate the best member method using the synthetic EPS system proposed by Wang and Bishop (2005), we varied K from 3 to 20 and tested values of 0.3 and 1.7 for μ_a, corresponding respectively to an underdispersive and an overdispersive EPS.

As proposed by Wang and Bishop (2005), we built each time a database of past forecasts by generating 15 000 observations y_t and corresponding ensemble forecasts $\mathbf{x}_t$. The different methods have then been compared using statistics computed on a second set of 15 000 ensemble forecasts $\mathbf{x}_t$. To obtain accurate statistics, N=150 statistical ensemble members have been drawn from each ensemble member $x_{t,k}$.

8.2 *A focus on variance and kurtosis*

It is clear from the experimental setup that the mean and skewness of both the observations and the forecasts will be zero, so we know before hand that the first and third moments of the forecasts will be reliable. We will therefore focus on the second and fourth moments, i.e. the variance and the kurtosis. While the variance measures the amount of dispersion of a distribution, the kurtosis, denoted by β_2 and defined by the ratio of the fourth central moment to the square of the variance, measures the degree of peakedness of a distribution. Distributions with high kurtosis values (above 3), are said to have heavy tails, because the probability mass located in the tails, far from the mean, is relatively high. We have shown by numerical integration that the kurtosis of the observations $\{y_t\}$ is very close to five, so that the distribution of the observations has quite heavy tails. If the kurtosis of the ensemble forecasts is not close to five, the probability of extreme events will not be correctly estimated on average. We will see that the method of Wang and Bishop (2005), while making the forecasts second-order reliable, leads to forecasts which have a kurtosis much larger than five, thus leading to an overestimation of the probability of extreme events.

8.3 *Results*

Figure 1(a) shows the ratio of the variance of the statistical ensemble members to the variance of the observations as a function of K, the dynamical ensemble size, for μ_a=0.3. Figure 1(b) shows the difference between the kurtosis of the ensemble members and the kurtosis of the observations as a function of K, also for μ_a=0.3. Figure 2(a) and 2(b) provide the same information, but for μ_a=1.7. The thick line on the graphs shows the ideal value, i.e. one for the variance ratio and zero for the kurtosis difference.

For an underdispersive EPS, it is clear from Figure 1(a) that both the method of Wang and Bishop (2005) and our method successfully adjust the variance of the ensemble forecasts so as to make them second-order reliable, but as Figure 1(b) shows, this leads to a very important increase in the kurtosis of the ensemble forecasts for the method of Wang and Bishop (2005). On the contrary, our method leads to a slight underestimation of the kurtosis, but still improves upon the original method of Roulston and Smith (2003). For an overdispersive EPS, Figure 2 shows that our method leads to second-order reliable forecasts, with a kurtosis still close to the kurtosis of the observations. Notice the change in scale between Figure 1(b) and Figure 2(b), meaning that the increase in kurtosis for the method of Wang and Bishop (2005) is much more important than for the method proposed in this paper. Note also that Figure 2 shows no results for the method of Wang and Bishop (2005), as this method is not applicable when the dynamical ensemble is already overdispersive, i.e. when μ_a>1.

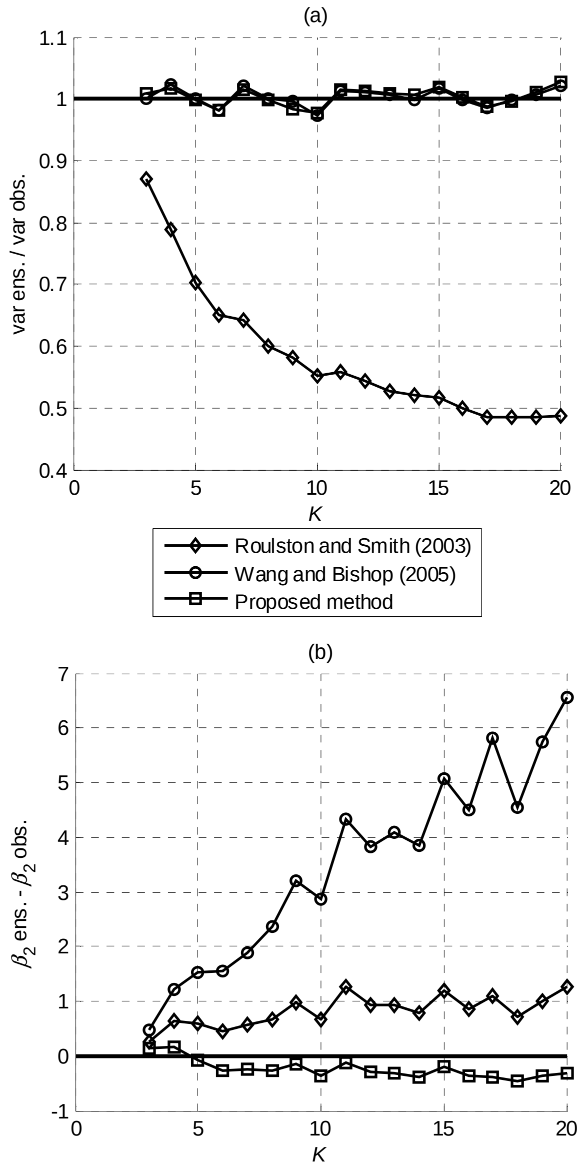

Figure 1: For μ_a=0.3, (a) Ratio of the variance of the statistical ensemble members to the variance of the observations and (b) difference between the kurtosis of the statistical ensemble members and the kurtosis of the observations

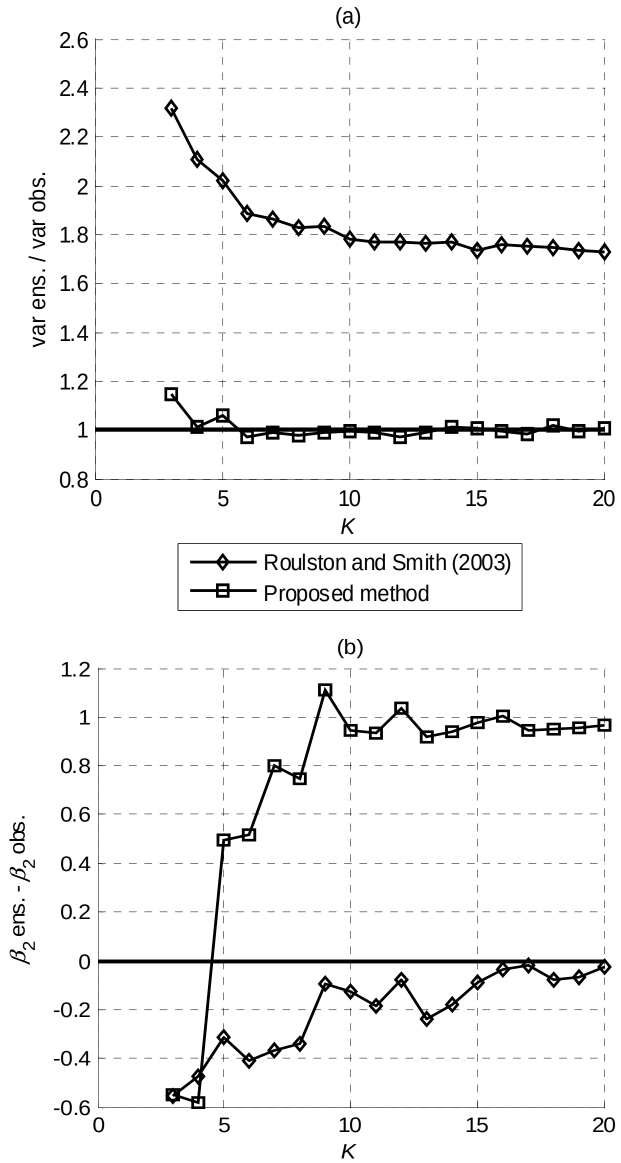

Figure 2: For μ_a=1.7, (a) Ratio of the variance of the statistical ensemble members to the variance of the observations and (b) difference between the kurtosis of the statistical ensemble members and the kurtosis of the observations

9. DISCUSSION AND CONCLUSION

Ensemble prediction systems (EPS) can be very useful to estimate the uncertainty of a deterministic numerical weather prediction. However, the number of members in an EPS is limited by computing resources and EPS outputs typically do not provide reliable probabilistic forecasts. We have shown in this paper that while the best member method of Roulston and Smith (2003), as modified by Wang and Bishop (2005), can lead to second-order reliable forecasts for an underdispersive EPS, it can also lead to probabilistic forecasts having very heavy tails, and which therefore will overestimate the probability of extreme events.

We have proposed and tested on a synthetic EPS a new method for dressing ensemble members which is based on the idea of dressing and weighing each ensemble member differently. This new method has the advantage of working both for underdispersive and overdispersive ensembles, and leads to forecasts that are more reliable, in the sense that they have enough variance, but also tail probabilities which are closer to those of the observations, especially for highly underdispersive ensembles.

As presented in this paper, the proposed method is only applicable to the problem of univariate forecasting, because the weight and error distribution of each ensemble member depends on its rank in the sample, but also because the best member of the ensemble is identified by the absolute difference between the observation and the forecast. If one wants to generalize the approach to multivariate forecasting, one possibility is simply to choose a norm in a multidimensional space which can be used both to identify the best member in the ensemble and to rank the ensemble members by their distance to the ensemble mean.

The major limitation of the method as proposed is the fact that it requires a database of historical forecasts from the best member of the ensemble in which each order statistic of the dynamical ensemble is well represented. When the size of the dynamical ensemble is large compared to the size of the historical database (which is often in practice quite small), it becomes difficult to estimate accurately the error distribution for each order statistic, as well as the weight that should be assigned to each order statistic.

Another difficulty, which is shared by both the original best member method of Roulston and Smith (2003) and the improved method of Wang and Bishop (2005), is the fact that the observed process and the ensemble must be stationary or deseasonalized prior to its analysis.

In the coming months, we plan to test the proposed method with outputs from the Canadian ensemble prediction system, with application to streamflow forecasting.

10. REFERENCES

Bernardo, J. and A. Smith, 1994: *Bayesian Theory*. Wiley.

Buizza, R., P.L. Houtekamer, Z. Toth, G. Pellerin, M. Wei and Y. Zhu, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble prediction systems. *Mon. Weather Rev.*, **133**, 1076-1097.

Fortin, V. and A.-C. Favre, 2005: Probabilistic forecasting from ensemble prediction systems: improving upon the best member method by using a different weight and dressing kernel for each member. *Q.J.R. Meteorol. Soc.*, submitted for publication.

Krzysztofowicz, R., 2004: Bayesian processor of output: a new technique for probabilistic weather forecasting. *Proc. of the 84th AMS Annual Meeting*, Seattle, 11-15 January 2004.

Raftery, A.E., T. Gneiting, F. Balabdaoui and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon. Weather Rev.*, **133**, 1155-1174.

Roulston, M.S. and L.A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A,** 16-30.

Sivillo, J.K., J.E. Ahlquist and Z. Toth, 1997: An ensemble forecasting primer, *Weather and Forecasting*, **12**, 809-818.

Wang, X. and C.H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteorol. Soc.*, **131**, 965-986.