J6.4 ENSEMBLE AUGMENTATION WITH A NEW DRESSING KERNEL

Xuguang Wang*
The Pennsylvania State University, University Park, PA
Craig H. Bishop
Naval Research Laboratory, Monterey, CA

1. INTRODUCTION

Forecast uncertainty is due to imperfect knowledge of the initial conditions and model errors. So far, most operational medium-range global ensemble forecasts (Molteni et al. 1996; Toth and Kalnay 1993, 1997) focus on the generation of initial perturbations running a single deterministic model. The deficiency of these ensembles is evident in the fact that observations often fall outside the range of ensemble values with a margin and frequency that cannot be explained by estimates of observational error, especially at longer forecast lead times. Besides yet imperfect ensemble representations of initial condition errors, the neglect of model errors (Smith 2001; Houtekamer et al. 1996) and the internal stochasticity is no doubt responsible for this non-ideal performance. The error in a single deterministic model inevitably exists due to its finite temporal and spatial resolution. While the developing of stochastic prediction models based on stochastic parameterizations of sub-grid scale processes could provide a first principle basis for accounting for model error, it is doubtful that such an approach would ever make full account of all types of errors. In other words, there are always residual errors missing from an ensemble.

To account for these residual errors, one way we can try is to add statistical perturbations to each ensemble member in the post-processing. This idea is first tried by the best member dressing technique by Roulston and Smith (2003). In the best member dressing method, the statistical perturbations are from archived historical best member errors. The best member is defined as the closest to the verification in the *full* space including all spatial locations, all quantities and all forecast lead times. Hereafter we call it the true best member. Our first concern about this best member dressing method is that since for an ensemble system, each member should have similar error statistics, one should not expect the errors of the first, the second and the even the worst member to be significantly different measured in the *full* space. Secondly, identification of the best member in the full space is time consuming. So one would choose a subspace to estimate which member is the true best member. The selection of the best member and thus the performance of the dressed ensemble are critically dependent on the choice of the subspace. Third, Roulston and Smith (2003) shows that using a too low dimensional space will very likely misidentify the true best member and it will underestimate the errors associated with each forecast. But there is no proof that the error statistics of the true best member or practically the best member identified from a high dimensional space will provide right ensemble spread, that is, no under-dispersion or over-dispersion. Our simple experiment with univariate random number generators reveals that the best member dressing method generally does not produce ensemble whose rank histogram is flat. It could be overdispersive or underdispersive depending on the undressed ensemble size and relationship of the undressed ensemble variance and the true ensemble mean error variance.

With these questions in mind we propose another dressing kernel in this paper. The basic idea is to choose the statistical perturbations that will make the dressed ensemble members indistinguishable from the verifications under the second moment measurements. The next section provides the mathematical expression for this dressing kernel. In sections 3 we describe the experiments we used to test this dressing kernel and reveal the problems of the best member method. In section 4 we show the comparison results measured with different tests. In section 5 we further discuss these results and in section 6 we summarize the results.

2. THE NEW DRESSING KERNEL

The distribution from which the imperfect ensemble is drawn is given by an infinite number of realizations of the stochastic process,

$$\mathbf{x}^f = \overline{\xi} + \xi', \qquad (1)$$

where $\mathbf{x}^f$ is a multidimensional random vector. $\overline{\xi}$ is its mean and ξ' is its deviation from its mean.

* Corresponding author address: Xuguang Wang, The Penn State University, University Park, PA, 16802. Email: xuguang@essc.psu.edu

The covariance of the ensemble distribution is given by

$$\boldsymbol{\Sigma}_e^2 = \left\langle \xi' \xi'^{\mathrm{T}} \right\rangle . \qquad (2)$$

The basic idea of the dressing technique is to add a statistical perturbation $\boldsymbol{\varepsilon}$ to the imperfect ensemble (1). Mathematically the dressed ensemble members are defined to draw from the stochastic process

$$\mathbf{y}^f = \bar{\xi} + \xi' + \boldsymbol{\varepsilon} . \qquad (3)$$

The basic question in the dressing technique then is how to choose $\boldsymbol{\varepsilon}$. In the best member method (Roulston and Smith 2003), $\boldsymbol{\varepsilon}$ is obtained from archived historical best member errors through resampling. In this paper we introduce a new method where $\boldsymbol{\varepsilon}$ is chosen aiming to make our dressed ensemble members indistinguishable from the verification, $\mathbf{z}^t$, under second moment measurement. Mathematically, we require

$$\left\langle \left(\mathbf{y}_i^f - \mathbf{y}_j^f\right)\left(\mathbf{y}_i^f - \mathbf{y}_j^f\right)^{\mathrm{T}} \right\rangle = \left\langle \left(\mathbf{y}_i^f - \mathbf{z}^t\right)\left(\mathbf{y}_i^f - \mathbf{z}^t\right)^{\mathrm{T}} \right\rangle , \qquad (4)$$

where $\mathbf{y}_i^f$ and $\mathbf{y}_j^f$ are two randomly picked dressed ensemble members, i.e.,

$$\mathbf{y}_i^f = \bar{\xi} + \xi_i' + \boldsymbol{\varepsilon}_i ,$$
$$\mathbf{y}_j^f = \bar{\xi} + \xi_j' + \boldsymbol{\varepsilon}_j . \qquad (5)$$

Substitute $\mathbf{y}_i^f$ and $\mathbf{y}_j^f$ from (5) into (4) and note that

$$\left\langle \boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_i^{\mathrm{T}} \right\rangle = \left\langle \boldsymbol{\varepsilon}_j \xi_i'^{\mathrm{T}} \right\rangle = \left\langle \xi_j' \xi_i'^{\mathrm{T}} \right\rangle = \left\langle \boldsymbol{\varepsilon}_i \xi_i'^{\mathrm{T}} \right\rangle = \mathbf{0} ,$$
$$\left\langle \left(\bar{\xi} - \mathbf{z}^t\right) \boldsymbol{\varepsilon}_i^{\mathrm{T}} \right\rangle = \left\langle \left(\bar{\xi} - \mathbf{z}^t\right) \xi_i'^{\mathrm{T}} \right\rangle = \mathbf{0} ,$$
$$\left\langle \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}} \right\rangle = \left\langle \boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_j^{\mathrm{T}} \right\rangle = \left\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \right\rangle ,$$
$$\left\langle \xi_i' \xi_i'^{\mathrm{T}} \right\rangle = \left\langle \xi_j' \xi_j'^{\mathrm{T}} \right\rangle = \boldsymbol{\Sigma}_e^2 . \qquad (6)$$

Then, we obtain

$$\left\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \right\rangle = \left\langle \left(\bar{\xi} - \mathbf{z}^t\right)\left(\bar{\xi} - \mathbf{z}^t\right)^{\mathrm{T}} \right\rangle - \boldsymbol{\Sigma}_e^2 . \qquad (7)$$

So if we obtain covariance of $\boldsymbol{\varepsilon}$ from (7) and parameterize a distribution, then we can use random number generator to generate the dressing perturbations. But for each individual forecast, there is only one realization of verification $\mathbf{z}^t$. So we have to relax our goal to make our dressed ensemble members indistinguishable from the verification on, for example, a seasonally averaged basis. In other words, we use forecasts and verifications from a season to calculate the first term on the right side of (7) and use the seasonally averaged ensemble covariance to replace the second term. Mathematically,

$$\left\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \right\rangle_s = \left\langle \left(\bar{\xi} - \mathbf{z}^t\right)\left(\bar{\xi} - \mathbf{z}^t\right)^{\mathrm{T}} \right\rangle_s - \left\langle \boldsymbol{\Sigma}_e^2 \right\rangle_s , \qquad (8)$$

where $\left\langle \cdot \right\rangle_s$ means seasonal average. Thus the statistics of the dressing perturbations are the same for all forecasts over a season. This property is the same as the best member method.

Note In the above analysis, $\bar{\xi}$ refers to the mean of the underlying distribution from which the ensemble is drawn. For finite ensemble sizes, it will not be equal to the mean of the ensemble. Similarly, $\boldsymbol{\Sigma}_e^2$ does not give the covariance of a finite ensemble about its mean, it gives the covariance of the distribution from which the ensemble was drawn. However it can be proved that for a configuration of our following experiment with a whole season's (3 months) 16-member ensemble runs, the two terms $\bar{\xi}$ and $\boldsymbol{\Sigma}_e^2$ can be approximated by the finite ensemble mean and finite ensemble covariance within 10% tolerance.

Note the covariance matrix given by (8), denoted as $\mathbf{Q}$ hereafter, is real and symmetric but not positive definite. We design the random generator in the following way. We first perform eigenvalue decomposition on $\mathbf{Q}$,

$$\mathbf{Q} = \mathbf{E}\boldsymbol{\Omega}\mathbf{E}^{\mathrm{T}} , \qquad (9)$$

where columns of $\mathbf{E}$ contain the eigenvectors and the diagonal matrix $\boldsymbol{\Omega}$ contains the corresponding eigenvalues. Positive eigenvalues indicate that on the directions of the corresponding eigenvectors the ensemble is underdispersive and thus dressing is necessary. Whereas for negative eigenvalues, on the corresponding eigenvectors the ensemble is overdispersive already and hence dressing is prohibited in these direction. Based on this argument, we define the random generator as

$$\boldsymbol{\varepsilon} = x_1 \mathbf{e}_1^+ + x_2 \mathbf{e}_2^+ + \cdots + x_k \mathbf{e}_k^+ , \qquad (10)$$

where $\mathbf{e}_i^+$, $i = 1 \cdots k$, are all eigenvectors corresponding to the positive eigenvalues. x_i, $i = 1 \cdots k$, are univariate random variables which are parameterized as normal distributions with mean equal to zero and variance equal to the ith positive eigenvalue of $\mathbf{Q}$, denoted as ω_i^+. So mathematically,

$$\left\langle x_i^{\ 2} \right\rangle = \omega_i^+ . \tag{11}$$

Note if all eigenvalues are positive and x_i, $i = 1 \cdots k$, are independent to each other, covariance of $\boldsymbol{\varepsilon}$ defined from (10) is equal to $\mathbf{Q}$ exactly for infinite samples.

Since the cost of generating an $\boldsymbol{\varepsilon}$ is that of generating a random vector of the same length, practically each member of the finite ensemble can be dressed with a very large number of $\boldsymbol{\varepsilon}$. In the following experiment, each dynamic ensemble member is dressed with the same number of $\boldsymbol{\varepsilon}$.

3. NUMERICAL EXPERIMENT

3.1 Ensemble System

To reveal the problems of the best member method, hereafter the RS method, and test our proposed dressing method, hereafter the WB method, we run 16-member spherical simplex ETKF ensemble with the NCAR community climate model (CCM3) initialized with NCEP/NCAR reanalysis. For details on the ETKF ensemble, please refer Bishop et al. (2001), Wang and Bishop (2003) and Wang et al. (2003). Different from the previous experiments where only the simulated rawinsonde observations are considered, simulated satellite observations are also included in the observational network in the current ETKF ensemble run.

3.2 Verifications

The verifications are NCEP/NCAR reanalysis located on the reanalysis grids that are nearest to the rawinsonde sites. The variables that we are interested in dressing and verifying are 500-hPa U over eastern USA for each individual forecast lead time. No temporal correlation is considered. 14 reanalysis grids over eastern USA are selected. The CCM3 ensemble outputs are interpolated to these grids. The training statistics of the bias and the dressing perturbations are from all 10-day ensemble runs of 1999 summer. Note before dressing the training bias is removed first. The verification period is summer 2001.

3.3 Experiment on the RS Method

In Roulston and Smith (2003) and Roulston (2003, personal communication), the RS method tries to estimate the true best member with limited number of variables. It is suggested (Roulston 2003, personal communication) that If practically feasible, the identification of the best member should be made using all quantities at all locations and all forecast lead times for which verifications are available even if the variables of interest are only in a small model subspace. In line of their argument, in our experiment, although we are only interested in 500-hPa U wind over eastern USA, we first identify the best member by using a quite high dimensional space, 500-hPa U over global verification sites throughout 1 to 10 day forecast lead times. Each sample of the best member error is stored in a vector. When dressing, as in Roulston and Smith (2003), each 1-10 forecast is dressed by one vector containing 1-10 day best member error.

To further reveal that in the RS method the dressing result is highly dependent on the way the best member is identified and to test whether the best member should be identified in a full or high dimensional space, we also try the experiment where the best member is defined in a relatively low dimensional space, 500-hPa U over eastern USA for each individual forecast lead time. That is, this subspace contains only those quantities of our interest. Forecasts from different forecast lead times are dressed separately.

Note the norm of the distance of an ensemble member and the verification is defined the same way as in equation (1) of Roulston and Smith (2003).

3.4 Experiment on the WB Method

Different form the RS method that has to decide which model subspace is chosen to build the statistics, in the WB method, the statistics of $\boldsymbol{\varepsilon}$ is built according to (8) just for the variables of interest, i.e., 500-hPa U over eastern USA (14 sites) for each individual forecast lead time. When dressing, random vectors of length 14 are generated by (10) and (11) for each forecast lead time separately.

4. COMPARISON RESULTS

In this section we compare the performance of the RS and WB dressing methods with different measurements. In these measurements, ensemble forecasts over the 14 sites for runs of 2001 summer are grouped together.

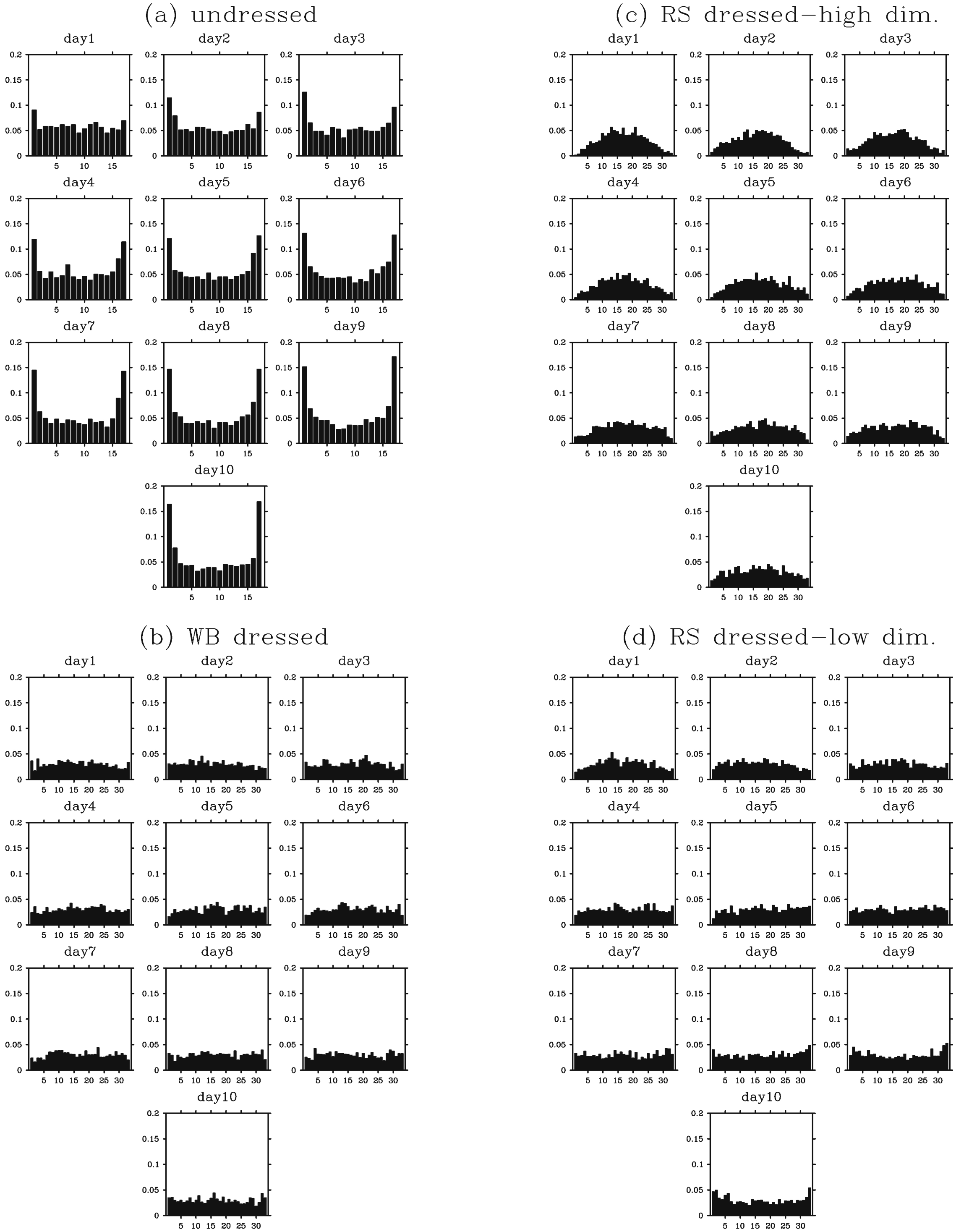

Fig. 1 Rank histograms for (a) the 16-member undressed ETKF ensemble, (b) 32-member WB-dressed ensemble, (c) 32-member RS-dressed ensemble with the best member identified from the high dimensional space defined in section 3.3 and (d) 32-member RS-dressed ensemble with the best member identified from the low dimensional space defined in section 3.3.

4.1 Rank Histogram

The first test is the rank histogram (Hamill 2001). Ensemble forecasts over the 14 sites for 2001 summer runs are grouped together. We dress each member of the 16-member undressed ETKF ensemble with 2 perturbations to form 32-member dressed ensembles. Figure 1(a) is the result for the undressed 16-member ensemble after removing the training bias. So, the undressed ensemble is underdispersive especially for longer forecast lead times, which is typical for a raw dynamic ensemble. The result for the WB-dressed ensemble is shown in fig. 1(b). The rank histogram for the WB method is flat throughout 1 to 10 forecast lead times, which indicates the WB dressing technique on average provides proper ensemble spread. Figure 1(c) is the result from the RS method where the best member is identified from the high dimensional space defined in section 3.3. The rank histogram indicates that through 1 to 10 day forecast lead times, this RS-dressed ensemble is overdispersive. For the RS method where the best member is identified from the low dimensional space defined in section 3.3, the result is better than that where the best member is identified from the high dimensional space. However it is still worse than the WB-dressed ensemble in that it is overdispersive at 1-2 day lead times and underdispersive at 8-10 day lead times (fig. 1d).

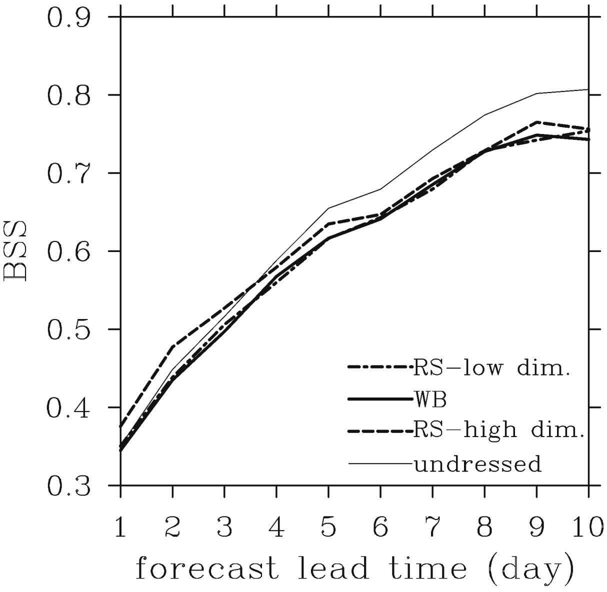

Fig. 2 Brier skill score (lower score is better) measurements for 1 to 10 day forecast lead times.

4.2 Skill Score

In fig. 2 we show the Brier skill score (Brier 1950; Murphy 1973) measurement results. In the calculation 4 climatologically equally likely bins are defined from 1999 summer verifications. We find the WB-dressed ensemble performs better (lower score is better) than the undressed ensemble throughout 1-10 day forecast lead times. The RS-dressed ensemble with the best member defined in the high dimensional space is inferior to the WB-dressed ensemble and at short lead times, it is even worse than the undressed ensemble. The result from the RS dressed ensemble with the best member defined in the low dimensional space is close to that from the WB method.

Another skill score we tried is the continuous ranked probability score (CRPS) by Hersbach (2000). The comparison result (not shown) is similar to that of the BSS.

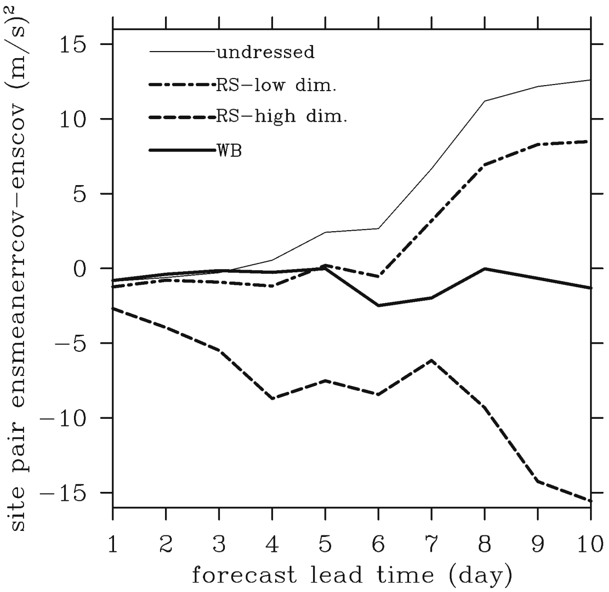

Fig. 3 Difference between the seasonally averaged ensemble mean error product of any two of the 14 sites and the seasonally averaged ensemble covariance of any two of the 14 sites from 1 to 10 day forecast lead times.

4.3 Ensemble covariance test

To measure the skill of the ensemble covariance, the first test we made is to see on average how different the ensemble predicted covariance between two sites different from the ensemble mean error covariance between the two sites. So for each forecast lead time, we first group respectively the ensemble covariance and the ensemble mean error product between any two of the 14 sites for the whole 2001 summer. Then we average the ensemble covariance and the ensemble mean error product respectively. The former gives the seasonally averaged ensemble covariance among two sites of the 14 sites and the latter approximates the corresponding seasonally averaged ensemble mean error covariance. For ensembles with accurate average prediction of the ensemble covariance, these two averaged values should be the same. In fig. 3, we plot the difference between these two averaged values for 1-10 day lead times. We find the result for the WB-dressed ensemble is fluctuating around zero throughout 1-10 day lead times. Whereas, the results from the undressed ensemble and the two RS-dressed ensembles deviate from zero largely especially at longer forecast lead times. Note the result of the RS-dressed ensemble with the best member defined from the high dimensional space has apparent deviations from zero even at the short lead times

Another test we made to measure the ensemble covariance is to see how the ensemble covariance can resolves the ensemble mean error covariance, i.e., the precision of the ensemble covariance. As in the above measurement, we first collect pairs of the ensemble covariance and the corresponding ensemble mean error product between any two of the 14 sites for the whole 2001 summer for each forecast lead time. Then we divide these points into 3 equally populated bins arranged in order of increasing ensemble covariance. Then we average the ensemble mean error product and the ensemble covariance respectively in each bin. For ensembles with fine precision of ensemble covariance, lines connecting these averaged points should have slope equal to 45 degrees and cross the point (0,0), hereafter refer to as the reference line. What is shown in fig. 4 is the averaged ensemble mean error product and the ensemble covariance for each bin for 1 day and 9 day forecast lead times (other lead times not shown for brevity). For the 1-day lead time, the line for the WB method is close to those of the undressed ensemble and the RS-dressed ensemble with the best member defined by the low dimensional space. They all have slope smaller than the reference line. The result from the RS-dressed ensemble with the best member defined by the high dimensional space deviates from the reference line much more than the other three ensembles. For the 9-day lead time, the result from the WB method is the closest to the reference line relative to all other ensembles.

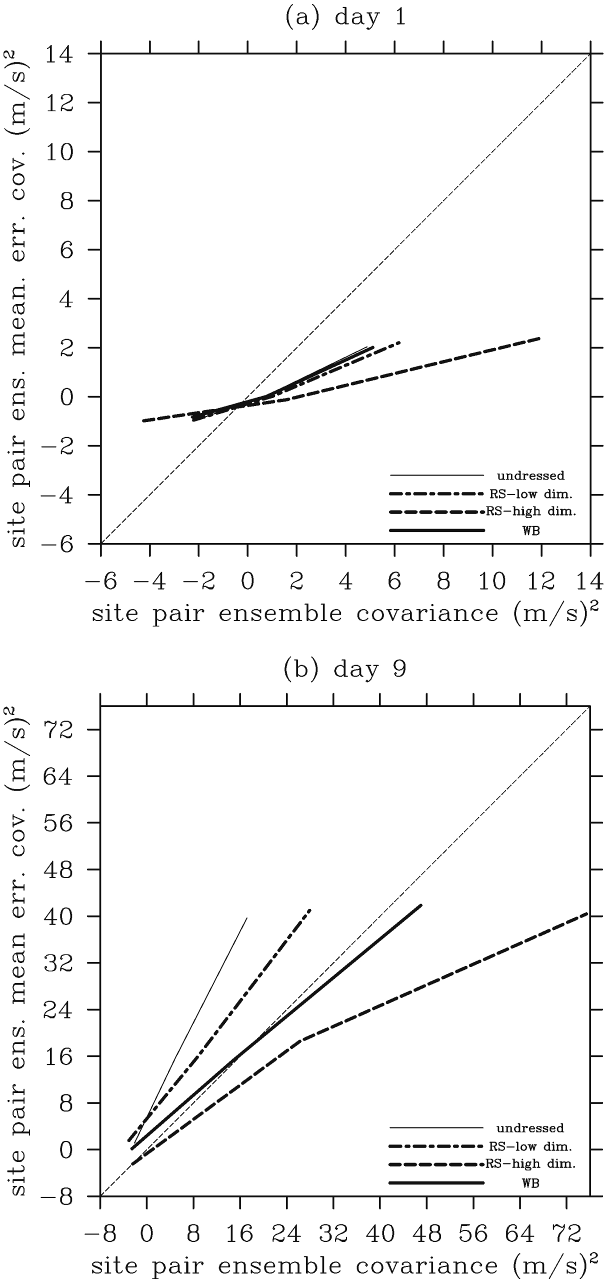

Fig. 4 Relationship between the averaged ensemble covariance and the averaged product of the ensemble mean error (approximation of the ensemble mean error covariance) of any two of the 14 sites for (a) 1-day forecast lead time and (b) 9-day forecast lead time. See details in section 4.3.

5 DISCUSSION ON THE RESULTS

The above comparison results verify our first concern about the RS method's basic hypothesis that the error statistics of the best member identified from the *full* space, i.e., the true best member, provides proper dressing perturbation error statistics. The higher the dimension used to identify the best member, the more likely the true best member is identified. Thus if the error statistics of the true best member indeed provides the proper statistics for the dressing perturbations, we should get better results using the high dimensional space defined in section 3.3 than the low dimensional space defined in section 3.3. However, the results in section 4 contradict this corollary. In the most measurements above, the RS-dressed ensemble with the best member defined in the high dimensional space is inferior to that with the best member defined in the low dimensional space. To explain why the RS ensemble with the best member defined in the high dimensional space is overdispersive throughout 1-10 day lead times (fig.1c), we first notice that the error variance of the best member defined in the high dimensional space as in section 3.3 is only 10% smaller than the worst member. In other words, all members can be regarded as "the worst" or "the best" if identified in such high dimensional space. There is no significantly best member if identified in the full or high dimensional space. This questions Roulston and Smith (2003)'s basic hypothesis, but consistent with the basic idea for ensemble construction, i.e., all members should have the same error statistics on average.

From section 4, the results of the RS-dressed ensemble highly depend on the choice of subspace where the best member is identified. This result reveals the uncertainty on how to choose the proper subspace to identify the best member.

Although the RS-dressed ensemble with the best member defined in the low dimensional space performs better than that with the best member defined in the high dimensional space, it still perform worse than the WB ensemble. The fact that it produces overdispersive ensemble at short lead times and underdispersive ensemble at long lead times (fig. 1d) indicates that the best member method in general does not provide reliable augmentation, which is consistent with the results of our simple random number generator experiment.

In contrast, the WB method has no requirement on determining a subspace in prior time. As long as one chooses the space of quantities of interest to build up the statistics, It will provide a reliable dressing. An simple example on its reliability is that when an ensemble is already overdispersive, the WB method will choose not to dress the ensemble while the RS method will still dress the ensemble to make it even more overdispersive.

6 SUMMARY AND FUTURE WORK

In this paper, we describe a new statistical dressing kernel to augment the dynamic ensemble in the post-processing. Different from the best member method by Roulston and Smith (2003) where the dressing perturbation statistics comes from the archived historic best member error, the new kernel is determined by making the dressed ensemble member indistinguishable from the verification under second moment measurement on a seasonally averaged basis.

We test this new dressing kernel and reveal the problem of the best member method with the ETKF ensemble (Bishop et al. 2001; Wang and Bishop 2003; Wang et al. 2003). In the test categories of rank histogram, skill scores and ensemble covariance precision, the new dressing kernel performs better than the best member method in general.

In our experiments, the CCM3 outputs are verified against the NCEP/NCAR reanalysis. In future work and operational usage we will use the real observation data. We will also try other distributions other than Gaussian distribution when generating the random vector by (9) and (10) to dress quantities whose errors are largely deviate from Gaussian distribution.

REFERENCES

Bishop, C. H., B. J. Etherton and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420-436.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

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

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

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

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

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

Smith, L. A., 2001: *Nonlinear dynamics and statistics (chapter 2)*, Alistair I. Mees (ed.), Birkhauser.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

—, and —, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**,1140-1158.

—, —, and S. J. Julier, 2003: Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev*., accept pending revision. Send email to xuguang@essc.psu.edu for the complete manuscript.