Q. J. R. Meteorol. Soc. (2005), **131**, pp. 965–986 doi: 10.1256/qj.04.120

Improvement of ensemble reliability with a new dressing kernel

By XUGUANG WANG[1]* and CRAIG H. BISHOP[2]

[1]*The Pennsylvania State University, PA, USA*
[2]*Naval Research Laboratory, Monterey, CA, USA*

(Received 6 August 2004, revised 15 December 2004)

SUMMARY

A new method of combining dynamical and statistical ensembles for the purpose of improving ensemble reliability for underdispersive ensembles is introduced. The method involves adding independent sets of N random four-dimensional 'dressing' perturbations to each of the K members of a dynamical ensemble forecast to obtain an $N \times K$ dressed ensemble. The new method mathematically constrains the stochastic process used to generate the statistical dressing perturbations so that it removes seasonally averaged errors in the second moment measures for originally underdispersive ensembles. A random-number generator experiment and an experiment with the ensemble transform Kalman filter (ETKF) ensemble generation scheme show that the previously proposed 'best-member' dressing method fails to reliably predict the second moment of the distribution of forecast errors, whereas the new dressing method reliably predicts this second moment. After being dressed with the second moment constraint method, the ETKF ensemble is more skilful than the undressed ensemble. The ETKF ensemble postprocessed with the new dressing method is applied for probabilistic forecasts of cooling degree-days (CDD) for Boston. It is shown that the new kernel's ability to account for temporally correlated forecast errors results in ensemble forecasts of CDDs with reliable spread, whereas the best-member method leads to an underdispersive ensemble of CDD forecasts.

KEYWORDS: Ensemble postprocessing Overdispersive Second moment measure Underdispersive

1. INTRODUCTION

During the last decade, ensemble forecasting has become an important part of numerical weather prediction (NWP). It has been operationally implemented for medium-range NWP (e.g. Toth and Kalnay 1993, 1997; Houtekamer *et al.* 1996; Molteni *et al.* 1996) and is being incorporated into short-range NWP (e.g. Du *et al.* 1997; Hamill and Colucci 1997, 1998; Stensrud *et al.* 1999; Hou *et al.* 2001; Grimit and Mass 2002; Stensrud and Yussouf 2003). Compared to a single deterministic forecast with high resolution, ensemble mean forecasts with relatively low resolution for each ensemble member can produce smaller root-mean-square errors. Moreover, ensemble forecasts can provide *flow-dependent* estimates of forecast errors, depicted by ensemble spread or expressed in forecast probabilities (e.g. Whitaker and Loughe 1998; Toth *et al.* 2001). Studies by Richardson (2000), Zhu *et al.* (2002), Palmer (2002) and Roulston *et al.* (2003) amongst others, have demonstrated that the economic value of ensemble forecasts is greater than that of a single deterministic forecast to a wide range of weather forecast users.

Managers of weather sensitive activities can benefit from probabilistic forecasts that reliably represent the probability distribution of the verifications given the ensemble forecast (e.g. Palmer 2002). However, because of the sub-optimal initial perturbation-generation techniques and the lack of consideration of model errors, rank histogram diagnostics show (e.g. Hamill and Colucci 1997, 1998) that outputs from raw ensembles may be biased and under-dispersive, which limits the predictive power of the ensemble. Hence, developing postprocessing methods to calibrate the outputs of ensemble forecasting systems has also been of interest. Methods include: reliability diagram statistics (e.g. Zhu *et al.* 1996; Krzysztofowicz and Sigrest 1999; Toth *et al.* 2001; Atger 2003); verification rank histogram statistics (Hamill and Colucci 1997, 1998; Eckel and

* Corresponding author, present address: NOAA Climate Diagnostic Center, 325 Broadway, R/CDC1, Boulder, CO 80305-3328, USA. e-mail: xuguang.wang@noaa.gov

Walters 1998); spread–skill relationship statistics (Atger 1999); smoothing with fitted probability distributions (Wilks 2002); Bayesian model averaging (Kass and Raftery 1995; Raftery *et al.* 2003); logistic regression techniques (Hamill *et al.* 2004); and the 'best-member dressing' method (Roulston and Smith 2003, hereafter RS03).

In this paper, we focus on developing a statistical method to reliably augment the spread of underdispersive ensembles, which is one of the problems for most current operational ensemble systems. The new method introduced is intended to improve upon RS03's dressing method, where statistical perturbations are added to each member of the dynamic ensemble in the postprocessing. The advantages of the dressing method are: (a) ensemble size can easily be increased as one can easily add many dressing perturbations to each member of the dynamic ensemble; (b) the dressing method can reflect some residual errors that the dynamic ensemble has not yet accounted for; (c) the dressing procedure maintains all information of the flow-dependent forecast uncertainty information in the dynamic ensemble; (d) the dressing method can easily be applied to calibrate ensemble outputs of multi-dimensional variables; and (e) the dressed-ensemble members can conveniently be applied to different types of user application functions.

In the best-member dressing method proposed by RS03, the best member out of each historical ensemble forecast is first identified, and the difference between the best member and the verification, i.e. the best-member error, is stored. The archive of the best-member errors is built from all historical ensemble forecasts available. When dressing, the statistical perturbations are drawn from the archived historical best-member errors. The best-member dressing perturbations are straightforward to construct. However, we notice that this approach does not guarantee the dressed ensemble to be 'reliable' (Wilks 1995). Specifically, a reliable ensemble should appear to be drawn from the same distribution as the verifying observations given the ensemble. The best-member dressing approach, however, does not mathematically constrain the distribution of the augmented (or 'dressed') ensemble to satisfy this condition under *any* measure. The purpose of this paper is to reveal the limitations of the best-member method, and to introduce a new dressing method to make the originally underdispersive ensemble become reliable after dressing under the second moment measure (for a one-dimensional verification, the second moment refers to variance; for a multi-dimensional verification, the second moment refers to covariance).

In sections 2 and 4 we explicitly demonstrate how the best-member dressing can result in distributions of the augmented ensemble being unreliable under second moment measures. In particular, we show that the best-member dressed ensemble may still be underdispersive or even become overdispersive, depending on, for example, the size of the undressed ensemble, how underdispersive the undressed ensemble is (section 2) and the subjective rules used to define the best member in practice (section 4). The prototype test in section 2 is based around ensembles generated with a random-number generator in which the difference between the distribution of undressed-ensemble members and the distribution of verifying observations can be controlled. The test in section 4 is based around an ensemble generated using the ensemble transform Kalman filter (ETKF; Bishop *et al.* 2001; Wang and Bishop 2003; Wang *et al.* 2004).

In section 3, we give the theoretical basis of the new second moment constrained dressing technique, and illustrate it using the ensemble generated with a random-number generator. In section 4, the performance of the new dressing technique is compared against the best-member dressing technique for improving the reliability of the 500 hPa zonal wind, U_{500}, ensemble forecasts from the ETKF ensemble made with the National Center for Atmospheric Research (NCAR) Community Climate Model Version 3 (CCM3, Jeffery *et al.* 1996). In section 5, both dressing techniques are further

tested by applying them to forecasts of 3-day accumulated cooling degree-days (CDDs) at Boston. The ability of both dressing methods to provide estimates of the covariance of multi-variables is demonstrated in this application. Concluding remarks follow in section 6.

2. Limitations of best-member dressing: the random-number generator experiment

In this section, we use a simple random-number generator experiment (for other examples of random-number generator experiments see e.g. Atger (2004)) to identify the limitations of the best-member dressing technique. Let us assume that for each case, a verifying observation y is drawn from a normal distribution with zero mean and standard deviation σ_t; in other words, assume that $y \sim N(0, \sigma_t)$. To simulate daily variation of σ_t^2, we let σ_t^2 be drawn from a chi-square distribution with a certain degree of freedom, df, represented by Chi(df). In the result shown below, $df = 3$. As a proxy for an underdispersive K-member ensemble, let us draw K random numbers x_k, $k = 1, 2, \ldots, K$, where each x_k represents a random draw from a normal distribution with a correct mean but an incorrect standard deviation σ_e, in other words we assume that $x_k \sim N(0, \sigma_e)$. To represent the underdispersion, we let $\sigma_e^2 = a\sigma_t^2$ where $0 < a < 1$. To further reflect that the underdispersion varies daily and the average underdispersion is different for different ensemble systems, we let a be a random number drawn from a uniform distribution from nine ranges, (0, 0.2), (0.1, 0.3), (0.2, 0.4), ..., (0.8, 1.0).

Below we describe 12 steps used to simulate the training, forecasting and verifying procedures for the best-member dressing for a given K and a given range of a. To build the training statistics for the best-member dressing perturbations we proceed as follows.

- Step 1: Draw a sample of σ_t^2 from Chi(3) and then draw a verification from $N(0, \sigma_t)$.
- Step 2: Draw a sample of a from a uniform distribution corresponding to one of the given ranges and then draw a K-member undressed ensemble from $N(0, \sigma_e)$, where $\sigma_e^2 = a \cdot \sigma_t^2$.
- Step 3: Find the ensemble member that is closest to the verification and find its distance from the verification.
- Step 4: Store this 'best-member error' in an archive.
- Step 5: Repeat steps 1–4 M times to obtain an archive of the M best-member errors for M cases and/or compute the sample variance σ_b^2 of the archive of the best-member errors.

Note that since we require that the undressed ensemble is drawn from a distribution with the same mean as the verifying observations, in this simplified case, the mean of the M archived best-member errors is zero when M approaches infinity. Having obtained this archive of errors, we then simulate the forecasting and dressing processes for a given K and a given range of a as follows.

- Step 6: repeat above steps 1 and 2 to generate a sample of verification and a K-member undressed ensemble.
- Step 7: Generate K independent N-member statistical ensembles of best-member errors either by randomly sampling from the archive or by drawing K independent sets of N random numbers ε_{kn}, $n = 1, 2, \ldots, N$; $k = 1, 2, \ldots, K$ where $\varepsilon_{kn} \sim N(0, \sigma_b)$.

- Step 8: The statistical ensembles are then combined with the dynamical ensemble to create an $N \times K$ member dressed ensemble ψ_{kn}, $k = 1, 2, \ldots, K$; $n = 1, 2, \ldots N$ using:

$$\psi_{kn} = x_k + \varepsilon_{kn}, \quad k = 1, 2, \ldots, K; n = 1, 2, \ldots, N, \tag{1}$$

  for each case.
- Step 9: Repeat steps 6–8 M' times to collect M' cases.

To verify the reliability of the dressed ensemble under the second moment measure, first note that if the verification were drawn from the same probability distribution as the ensemble, then the average square distance between any two randomly selected dressed-ensemble members ought to be the same as the average square distance between randomly selected ensemble members and the verification. Consequently, we test whether the best-member dressing results in a reliable ensemble as follows.

- Step 10: Compute the averaged square distance between each distinct pair of dressed-ensemble members. Note that since the total number of dressing perturbations is different from the number of undressed-ensemble members, from Eq. (1) this quantity is calculated by $Term1 = \langle\langle(x_{mk} - x_{mi})^2\rangle_{i\neq k}\rangle_m + \langle\langle(\varepsilon_{mkn} - \varepsilon_{mil})^2\rangle_{kn\neq il}\rangle_m$, where subscript m denotes the mth case of the M' cases, $\langle\ \rangle_{i\neq k}$ is the average over all combinations of distinct undressed-ensemble members for the mth case, $\langle\ \rangle_{kn\neq il}$ is the average over all combinations of distinct dressing perturbations for the mth case, and $\langle\ \rangle_m$ is the average over all M' cases.
- Step 11: Compute the mean square distance between the verifying observations and each ensemble member by $Term2 = \langle\langle(\psi_{mkn} - y_m)^2\rangle_{kn}\rangle_m$ where $\langle\ \rangle_{kn}$ is the average over all dressed-ensemble members for the mth case.
- Step 12: Compare the relative difference (denoted by $DIFF$) of the quantities in steps 10 and 11, i.e. calculate $DIFF = (Term1 - Term2)/Term2$.

Steps 1 to 12 are repeated for different choices of K and range a.

Figure 1(a) shows $DIFF$ as a function of K and the average of $a = \sigma_e^2/\sigma_t^2$ for each range, i.e. $\overline{a} = 0.1, \ldots, 0.9$. Here, $M = M' = 15\,000$, and $N = 150$. Negative (positive) $DIFF$ indicates that the dressed ensemble is under-dispersive (over-dispersive). Figure 1(a) shows that for $K = 1$, $DIFF$ is equal to zero for all $\overline{a}$. When K is larger than 1, for any given $\overline{a}$, there is only one value of K that renders the best-member dressing method reliable. The best-member dressed ensemble is either overdispersive or underdispersive beyond that regime, depending on the undressed-ensemble size K and how under-dispersive (measured in $\overline{a}$) the undressed ensemble is. A contour plot (not shown) of the ratio of averaged dressed-ensemble variance over averaged true variance has a similar pattern to Fig. 1(a), that is, for a given $\overline{a}$, the ratio varies depending on K. For example, for $\overline{a} = 0.7$, although the dressed ensemble becomes less underdispersive as the ratio is greater than 0.7 for all K considered, the ratio is greater than 1 for $K < 5$, which means the dressed ensemble becomes overdispersive and the ratio is less than 1 for $K > 5$, which means the dressed ensemble is still underdispersive. Also note, if we draw perturbations directly from the archive rather than from a prescribed distribution, the best-member method has the limitation that the number of dressing perturbations is constrained by the size of the archive, M, because one needs to draw independent perturbations for different ensemble member. In the next section we introduce a new dressing kernel that does not suffer from the above limitations.

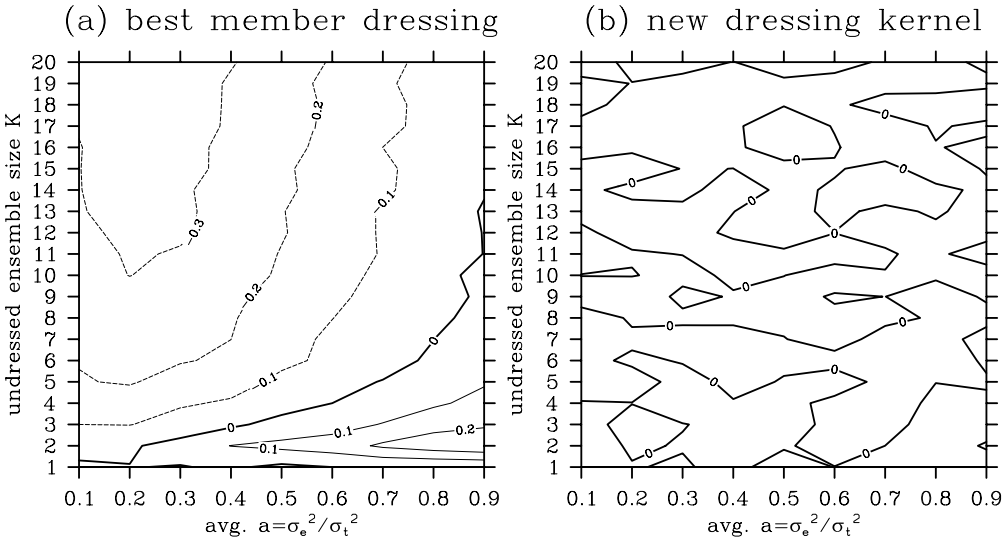

Figure 1. Random-number-generator experiment results in testing the reliability of the spread of the ensemble dressed by: (a) the best-member method, and (b) the new dressing kernel. Thin solid contours indicate overdispersive ensemble; dashed contours indicate underdispersive ensemble; thick solid contours mean the spread is reliable. See text for details.

3. DRESSING WITH THE SECOND MOMENT CONSTRAINT

We seek a mathematical constraint on a dressing kernel that will render the ensemble reliable on the seasonally averaged second moment measure. The method is applicable for underdispersive ensembles, which is one of the problems in current operational ensembles. For each case of forecasts over a season, let $\mathbf{y}$ contain a list of verifications that we wish to predict, and let $\mathbf{x}$ contain the corresponding list of forecast variables from one member of the underdispersive dynamic ensemble. To begin, we remove the seasonally averaged bias of each ensemble member and then assume that each ensemble member is drawn from a stochastic process:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{x}', \tag{2}$$

where $\langle \mathbf{x} \rangle = \overline{\mathbf{x}}$ gives the mean of the underlying distribution (which may be different to the sample mean one obtains when one takes the mean of a K-member ensemble) and $\langle \mathbf{x}' \rangle = 0$. The covariance of Eq. (2) is denoted as:

$$\mathbf{\Sigma}^2 = \langle \mathbf{x}' \mathbf{x}'^{\mathrm{T}} \rangle, \tag{3}$$

where T denotes the transpose. To dress the ensemble, statistical perturbations $\boldsymbol{\varepsilon}$ are added to each dynamic ensemble member. Let $\boldsymbol{\psi}$ list the corresponding dressed forecasts. Written in a similar format to Eq. (2), the dressed-ensemble members are drawn from the stochastic process:

$$\boldsymbol{\psi} = \mathbf{x} + \boldsymbol{\varepsilon} = \overline{\mathbf{x}} + \mathbf{x}' + \boldsymbol{\varepsilon}, \tag{4}$$

where $\langle \boldsymbol{\varepsilon} \rangle = 0$, $\langle \boldsymbol{\varepsilon} \mathbf{x}' \rangle = 0$. Note that the mean of the dressed ensemble is still $\overline{\mathbf{x}}$. Also note that we have assumed the seasonally averaged bias of $\overline{\mathbf{x}}$ has been removed. The basic idea of the new dressing kernel is to choose the covariance of $\boldsymbol{\varepsilon}$, that is $\langle \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\mathrm{T}} \rangle$, to make $\boldsymbol{\psi}$ indistinguishable from the verification, $\mathbf{y}$, under second moment measurement

on a seasonally averaged basis. Mathematically, we require:

$$\langle\langle(\boldsymbol{\psi}_{li}-\boldsymbol{\psi}_{lj})(\boldsymbol{\psi}_{li}-\boldsymbol{\psi}_{lj})^{\mathrm{T}}\rangle_{i\neq j}\rangle_l=\langle\langle(\boldsymbol{\psi}_{li}-\mathbf{y}_l)(\boldsymbol{\psi}_{li}-\mathbf{y}_l)^{\mathrm{T}}\rangle_i\rangle_l, \tag{5}$$

where subscript l denotes the lth case over a season, and subscripts i and j denote any two different dressed-ensemble members; $\langle\cdot\rangle_l$ represents the average of all cases over a season, $\langle\cdot\rangle_{i\neq j}$ denotes averaging over all combinations of any two different dressed-ensemble members for the lth case, and $\langle\cdot\rangle_i$ is the averaging over all choices of i for a particular case. Substituting Eqs. (3) and (4) into Eq. (5), one can show (see appendix) that Eq. (5) is satisfied provided that:

$$\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle=\langle(\overline{\mathbf{x}}_l-\mathbf{y}_l)(\overline{\mathbf{x}}_l-\mathbf{y}_l)^{\mathrm{T}}\rangle_l-\langle\boldsymbol{\Sigma}_l^2\rangle_l, \tag{6}$$

where $\overline{\mathbf{x}}_l$ and $\boldsymbol{\Sigma}_l^2$ are the mean and covariance of the *underlying* distribution from which the undressed ensemble is drawn for the lth case. Note that the covariance of the dressing perturbations $\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle$ is the same for all ensemble members for all cases; therefore, we put no subscript on this term. Also note that for a one-dimensional verification, Eq. (6) simply states that $\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle$ should be equal to the difference between the seasonally averaged variance of the error of the underlying ensemble mean and the seasonally averaged ensemble variance.

To understand the new dressing kernel $\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle$ given by Eq. (6), we use a two-dimensional figure (Fig. 2) to illustrate the idea. Assume the number of variables that we are interested in forecasting is two, that is, $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\psi}$ contain two elements each. Each black dot in Fig. 2(a) represents the difference between one ensemble forecast member and the corresponding underlying ensemble mean. Since there are L forecasts each of K members made each season, the number of dots present in Fig. 2(a) is equal to $K\times L$, and the covariance of these points corresponds to the $\langle\boldsymbol{\Sigma}_l^2\rangle_l$ term in Eq. (6). The one-sigma ellipse associated with this covariance is shown by the black line in Fig. 2(a). In Fig. 2(b), each of the L grey dots gives the difference between a verification and a corresponding underlying ensemble mean. The covariance of these dots gives the first right-hand term in Eq. (6). Since the seasonally averaged bias of undressed ensembles has been removed, the grey dots in Fig. 2(b) centre at (0, 0). Note that the one-sigma ellipse for the grey dots is larger than that for the black dots, indicating that the undressed ensemble in Fig. 2(a) is under-dispersive. In Fig. 2(c) we illustrate what Fig. 2(a) will look like after we dress one ensemble member with a number of perturbations. After we dress all members, the corresponding plot is shown in Fig. 2(d) where the scattered stars are the differences of the dressed-ensemble members from the corresponding underlying ensemble mean. The one-sigma ellipse associated with the stars is also shown by the black line in Fig. 2(d). The idea of the new dressing kernel is to constrain the covariance of the dressing perturbations, i.e. the one-sigma ellipse in Fig. 2(c), by Eq. (6), so that the one-sigma ellipse associated with the dressed-ensemble perturbations (stars) in Fig. 2(d) is identical to the one-sigma ellipse associated with the errors of the underlying ensemble mean (grey dots) in Fig. 2(b). Note that in Fig. 2(d) $\langle(\boldsymbol{\Sigma}^D)_l^2\rangle_l$ denotes the seasonally averaged covariance of the dressed-ensemble perturbations.

To obtain the underlying ensemble mean and covariance in Eq. (6), one would need an infinitely large ensemble. For a finite undressed-ensemble size, the *underlying* ensemble mean and covariance $\overline{\mathbf{x}}_l$ and $\boldsymbol{\Sigma}_l^2$ in Eq. (6) are estimated using a *sample* ensemble mean $\overline{\mathbf{x}}_l^s$ and a sample ensemble covariance $\boldsymbol{\Sigma}_l^{s^2}$, namely:

$$\overline{\mathbf{x}}_l^s=\frac{1}{K}\sum_{m=1}^{m=K}\mathbf{x}_{lm}, \tag{7}$$

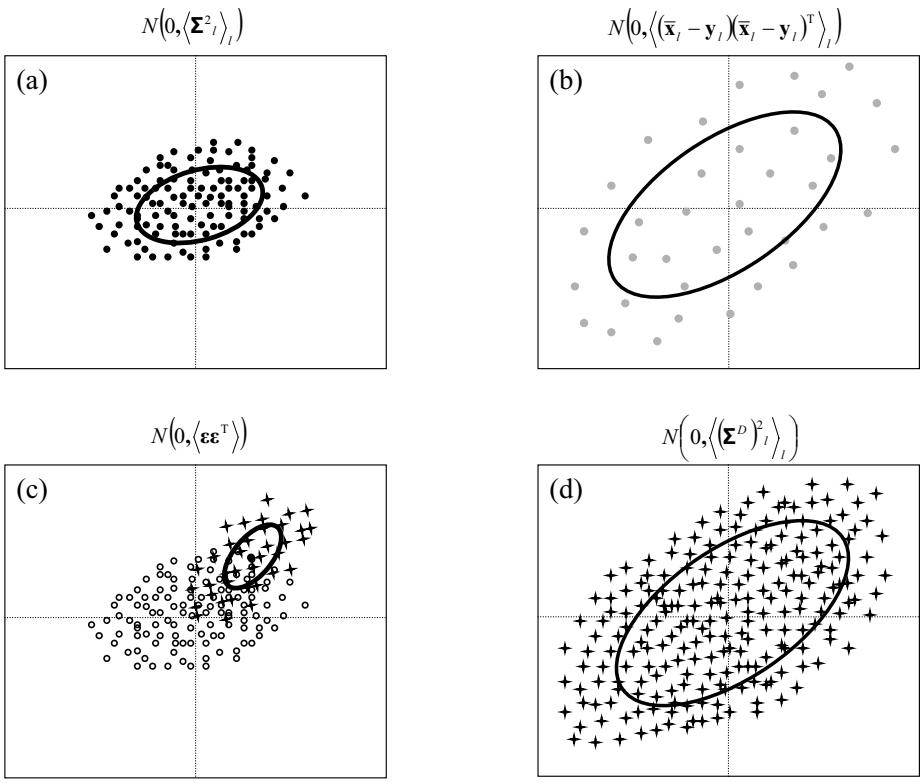

Figure 2. Illustration showing the idea of the new dressing kernel in two-dimensional space. (a) Black dots each represent the difference between one ensemble forecast member and the corresponding underlying ensemble mean; (b) grey dots each give the difference between a verification and a corresponding underlying ensemble mean; (c) shows (a) after one ensemble member is dressed with a number of perturbations; (d) scattered stars are the differences between the dressed-ensemble members from the underlying ensemble mean. The black lines are 1-sigma ellipses associated with the covariance of the corresponding points in each panel. See section 3 for a more detailed explanation.

and

$$\mathbf{\Sigma}_{l}^{s^2}=\frac{1}{(K-1)}\sum_{m=1}^{m=K}(\mathbf{x}_{lm}-\overline{\mathbf{x}}_{l}^{s})(\mathbf{x}_{lm}-\overline{\mathbf{x}}_{l}^{s})^{\mathrm{T}}. \tag{8}$$

Recall that seasonally averaged biases of the sample ensemble means are assumed to be removed from Eqs. (7) and (8). When sample covariances and means are used to estimate the terms in Eq. (6), the variation of the sample mean about the underlying ensemble mean must be accounted for. In the appendix, it is shown that accounting for this variation leads to the formula:

$$\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle=\langle(\overline{\mathbf{x}}_{l}^{s}-\mathbf{y}_{l})(\overline{\mathbf{x}}_{l}^{s}-\mathbf{y}_{l})^{\mathrm{T}}\rangle_{l}-\left(1+\frac{1}{K}\right)\langle\mathbf{\Sigma}_{l}^{s^2}\rangle_{l},\quad\text{for } K\geqslant 2. \tag{9a}$$

In the situation wherein there is only one control forecast $\mathbf{x}_{l}^{\mathrm{c}}$ for the lth case, that is $K=1$, the new dressing kernel is:

$$\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle=\langle(\mathbf{x}_{l}^{\mathrm{c}}-\mathbf{y}_{l})(\mathbf{x}_{l}^{\mathrm{c}}-\mathbf{y}_{l})^{\mathrm{T}}\rangle_{l},\quad\text{for } K=1. \tag{9b}$$

To test the new dressing kernel, we also adopt the one-dimensional random-number generator experiment in section 2 with the one-dimensional new dressing kernel given

by Eqs. (9a) and (9b). The result is shown in Fig. 1(b), which demonstrates that the new dressing kernel of Eqs. (9a) and (9b) can provide a reliable ensemble for all K and $\overline{a}$ under the second moment measure given by steps 10–12 in section 2.

To generate dressing perturbations for multi-dimensional variables, we use a multi-dimensional random-number generator. First, note that the covariance matrix given by Eq. (9), denoted as $\mathbf{Q}$ hereafter, is real and symmetric but not necessarily positive definite. The new dressing procedure applies dressing perturbations only in phase-space directions that show underdispersion. We first perform an eigenvalue decomposition on $\mathbf{Q}$:

$$\mathbf{Q} = \langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \rangle = \mathbf{E}\boldsymbol{\Omega}\mathbf{E}^{\mathrm{T}}, \tag{10}$$

where the 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. In contrast, negative eigenvalues indicate that the undressed ensemble is overdispersive in the directions of the corresponding eigenvectors. Since dressing the ensemble in the overdispersive directions would make it even more overdispersive in these directions, we only dress in the directions corresponding to positive eigenvalues. Similarly, directions of zero eigenvalues need not be dressed. Note that the best-member method does not have constraints not to dress the raw ensemble in these directions. Based on this argument, we define the new dressing perturbation generator as:

$$\boldsymbol{\varepsilon} = x_1\mathbf{e}_1^+ + x_2\mathbf{e}_2^+ + \cdots + x_I\mathbf{e}_I^+, \tag{11}$$

where $\mathbf{e}_i^+$, $i = 1, 2, \ldots, I$, are all eigenvectors corresponding to the positive eigenvalues. The coefficients x_i, $i = 1, 2, \ldots, I$, are univariate random variables which are parametrized as normal distributions with mean equal to zero and variance equal to the ith positive eigenvalue of $\mathbf{Q}$, denoted as ω_i^+. Mathematically:

$$x_i \sim N(0, \sqrt{\omega_i^+}). \tag{12}$$

Note that Eqs. (11) and (12) enable the generation of multi-dimensional dressing perturbations for the multivariate verification of interest at small cost. Also note that the new dressing kernel is designed for underdispersive ensemble. It is not only able to make underdispersive ensembles, after dressing, have reliable spread for each individual variable, but can also produce a reliable estimate of the error covariance between the variables of interest if all phase-space directions of the undressed ensemble show underdispersion. Depending on the variables of interest, the new dressing kernel can be constructed to consider both temporal and spatial correlations of the forecast errors. Thus, the method allows four-dimensional dressing. The new dressing perturbations can also be drawn from an archive instead of a prescribed distribution. The method by which this can be done is discussed in section 6.

Note that when K is greater than unity but rather limited, the new kernel defined by Eq. (9a) makes the dressed ensemble satisfy the condition that the seasonally averaged covariance of the differences between ensemble members and the verifications is equal to the seasonally averaged covariance of the differences between ensemble members. This is a useful second moment property. Another useful but slightly different second moment property, is to make the seasonally averaged covariance of the differences of the ensemble from the *sample* ensemble mean equal to the seasonally averaged error covariance of the sample ensemble mean. This latter property can be obtained

by replacing Eq. (9a) with:

$$\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \rangle = \langle (\overline{\mathbf{x}}_l^s - \mathbf{y}_l)(\overline{\mathbf{x}}_l^s - \mathbf{y}_l)^{\mathrm{T}} \rangle_l - \left(1 - \frac{1}{K}\right) \langle \boldsymbol{\Sigma}_l^{s^2} \rangle_l, \quad \text{for } K \geqslant 2. \tag{13}$$

In other words, from Eqs. (9a) and (13), these two second-moment properties cannot be satisfied simultaneously for smallish K. However, as K tends to infinity both properties are simultaneously satisfied. When one's forecast application relies solely on the ensemble mean, using Eq. (13) to define the new kernel is probably the best option. In contrast, when one's forecast application relies on a forecast probabilistic distribution, using Eq. (9a) to define the new kernel is probably the best option. The random-number generator experiment (not shown) demonstrates that when $K > 10$ one of the two properties can be satisfied precisely, and the other can be satisfied approximately by the new dressing kernel either defined by Eq. (9a) or Eq. (13). The best-member dressing kernel, however, does not satisfy either second moment property. Since in the ETKF ensemble experiments (described in section 4) $K = 16$, the results obtained with Eq. (9a) are very similar to those obtained with Eq. (13).

4. Test with a nonlinear CCM3 ETKF ensemble

The best-member dressing method was first designed and tested by RS03 with the nonlinear ensemble prediction system of the European Centre for Medium Range Weather Forecasts. The error statistics of nonlinear systems on a given day are usually non-Gaussian. In the random-number generator experiment of section 2 and 3 we assume a Gaussian error system. To check the performance of the best-member dressing and the new dressing methods in the nonlinear system with non-Gaussian error statistics, we apply both dressing methods to the 1- to 10-day CCM3 ETKF nonlinear atmospheric ensemble forecasts (Bishop *et al.* 2001; Wang and Bishop 2003; Wang *et al.* 2004). We also use this section to illustrate the sensitivity of best-member dressing to the manner in which one defines the 'best ensemble member'.

(*a*) *Numerical experiment design*

(i) *Dynamic ensemble, verification data, and variables of interest.* The ensemble to be dressed is a 16-member spherical simplex ETKF ensemble throughout 1- to 10-day forecasts. The ensemble is run on the NCAR CCM3 (Jeffery *et al.* 1996) and the initial conditions for each control forecast are obtained from the National Centers for Environmental Prediction (NCEP)/NCAR re-analysis (Kalnay *et al.* 1996). The observational network in the current experiment simulates both rawinsonde and satellite observations. For details on the construction of the spherical simplex ETKF ensemble, please refer to previous experiments in Wang and Bishop (2003) and Wang *et al.* (2004).

The verifications are NCEP/NCAR re-analysis data located on the re-analysis grids that are nearest to known rawinsonde sites. The variable we are interested in dressing is U_{500} over 14 re-analysis grids over the eastern USA (Fig. 3) at individual forecast lead times. The CCM3 ensemble outputs are interpolated to these grids during the training and validating phases of the experiment.

(ii) *Identification of the best member.* In RS03 the normalized distance between the ith ensemble member $\mathbf{x}_i$ and the verification $\mathbf{y}$ in the space of d variables is defined as:

$$R_{i,d}^2 = \sum_{k=1}^{d} \frac{(x_{i,k} - y_k)^2}{\Omega_k^2}, \tag{14}$$

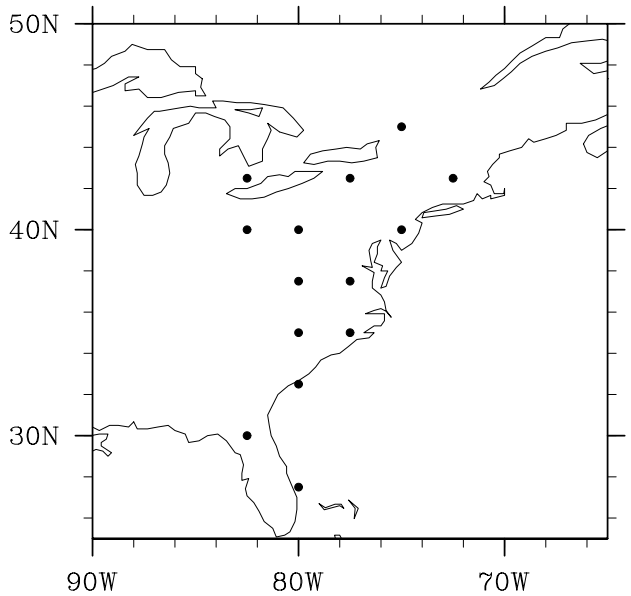

Figure 3. Locations of the 14 verification sites over the eastern USA used for the experiment in section 4.

where Ω_k^2 is the variance of $x_{i,k}$, the kth component of $\mathbf{x}_i$, and y_k is the kth component of $\mathbf{y}$. In RS03, the best member is defined as that closest to the verification in the *full* space, i.e. the space of d variables includes all spatial locations, all quantities and all forecast lead times. However, using the full space to make the identification is time consuming. RS03 tried to empirically determine the minimum number of variables that are unlikely to lead to misidentification. They suggested that, if practically feasible, high-dimensional space should be used even if the variables that we are interested in dressing are only in a small subspace. To test whether the best-member error statistics with the best member identified in a high-dimensional space can provide reliable spread, we use a quite high-dimensional space U_{500} over global verification sites throughout 1- to 10-day forecast lead times to identify the best member, although we are only interested in U_{500} over the 14 sites for each individual lead time. This subspace for identifying the best member is denoted as RS-10d-globe.

To reveal that the spread of the best-member dressed ensemble may not be reliable due to the uncertainty in selecting the subspace to identify the best member, we also try experiments where the best member is defined in two relatively low-dimensional spaces. One is U_{500} over the 14 eastern USA sites for each individual verification lead time, denoted as RS-id-east; the other is U_{500} over the 14 sites from day 1 till the verification lead time, denoted as RS-1-id-east.

(iii) *Training and forecasting processes.* The training statistics for bias and dressing perturbations are obtained from forecasts during the summer (June, July and August) of 1999. The U_{500} bias is obtained for each verification site for each forecast lead time by averaging the corresponding ensemble mean errors collected from 16-member ETKF runs during the 1999 summer. Before generating training statistics for the dressing perturbations for both the new kernel and the best-member method, the bias is first removed from each member of the 16-member ETKF ensemble for each verification site and at lead times of 1, 2, . . . , 10-days.

Since we are interested in U_{500} forecasts over the 14 verification sites at individual lead times, the new dressing kernel is constructed for each forecast lead time independently. In Eq. (9), vector $\overline{\mathbf{x}}_l^s$ contains 14 elements corresponding to the 500 hPa ensemble mean U_{500} forecasts at the 14 sites of the lth case during 1999 summer for each particular lead time. Vector $\mathbf{y}_l$ contains the corresponding verifications and $\mathbf{\Sigma}_l^{s^2}$ is the 14×14 ensemble covariance matrix. The resultant $\mathbf{Q}$ matrix is 14×14.

For the best-member method, the best member out of each 16-member ETKF run during the 1999 summer is selected first for the three subspaces. For the subspaces RS-id-east and RS-1-id-east, the best-member errors corresponding to U_{500} over the 14 verification sites are stored in a vector of 14 elements for each lead time. The archive of the best-member errors is built by archiving these vectors for each lead time over all runs of the 1999 summer. For the subspace RS-10d-globe, the index of the ensemble member that is the best member identified in the subspace of RS-10d-globe is the same for all lead times. In this case, the best-member errors are stored in a vector of 140 elements for each 1- to 10-day run. The first 14 elements store the errors of the best member over the 14 sites for a 1-day lead time, and the second 14 elements store the errors of the same member for 2-day lead time, and so on. The archive of the best-member errors for RS-10d-globe is then built by collecting such vectors from all 10-day forecasts over the 1999 summer.

To perform an out-of-sample test of the dressing techniques, forecasts were made for the 2001 northern hemisphere summer. For each 16-member ETKF run during 2001 summer, the training bias is first removed from each ensemble member. Independently sampled dressing perturbations are then added to each of the 16 members. For the new dressing kernel, 14-dimensional vectors are randomly generated using Eqs. (10)–(12) for each forecast lead time, and added to each member of the 16-member U_{500} forecasts over the 14 sites. For RS-id-east and RS-1-id-east methods, random 14-dimensional vectors are randomly drawn from the corresponding archives for each forecast lead time. For the RS-10d-globe method, random vectors of length 140 are randomly drawn from the corresponding best-member error archive. As mentioned above, the 140 elements contain ten sets of 14-dimensional vectors corresponding to lead times of 1 to 10 days. The first set of 14 elements is added to the ensemble forecast over the 14 verification sites for the 1-day lead time, the second set is added to the same ensemble forecast for the 2-day lead time, and so on.

(*b*) *Experiment results*

The performances of the dressed ensembles are measured by the rank histogram and probability scores. For each forecast lead time, samples are collected from all ensemble forecasts during the 2001 summer over the 14 verification sites. For the best-member method, if the dressing perturbations are drawn from the best-member error archive, the number of dressing perturbations to be added to each ETKF ensemble member is limited by the length of the time period during which the best-member error is collected. As we built the archive from one season's forecasts, the number of best-member dressing perturbations is limited by O(10) in order to make the dressing perturbations for each of the 16 ETKF ensemble members diverse enough. On the other hand, we want the number of dressing perturbations to be large enough so that the probability distribution derived from the dressed ensemble will be smooth, and also so that the ensemble mean whose seasonal average bias is removed will not be shifted due to the sampling error of the dressing perturbations. In our experiment we tried to dress each member of the 16-member ETKF ensemble with 2, 8, 16, and 32 perturbations. We found that the results start to converge when the number of dressing perturbations approaches 16 and 32.

The latter renders the sampling error of the dressing perturbation mean to be less than 5%. In the results shown in this section, each member of the 16-member ETKF ensemble has been dressed with 32 perturbations, thus yielding 512-member dressed ensembles. For the best-member method, the 32 perturbations are drawn from the best-member error archive. For the new dressing kernel, the 32 perturbations are drawn from multi-dimensional Gaussian distribution following Eqs. (10)–(12).

The first measurement of the reliability of the ensembles is applicable to scalar verifications and is called the rank histogram (Anderson 1996; Hamill 2001). Recall that the sizes of the undressed and the dressed ensembles are 16 and 512, respectively. Because the number of verifications to construct the rank histogram is limited relative to the rank of 512, and also because we want the y-axis of the histogram to have the same scale for the dressed and undressed ensembles, instead of constructing the histogram for the dressed ensemble by using all 512 dressed members we randomly choose 16 out of 512 members for each sample. Figure 4(a) is the result for the undressed 16-member ensemble for the 2001 summer after removing the bias from the 1999 summer (only results for days 1, 3, 5, 7, and 9 are shown in Fig. 4). The undressed ensemble is under-dispersive, especially for longer forecast lead times. The χ^2 test for the uniformity of the rank histogram (Wilks 1995; Anderson 1996; Hamill 2001) rejects the null hypothesis that the rank histogram is flat with a confidence level much higher than 99% (the P value is equal to 7.1×10^{-4} for day 1 and much smaller than 10^{-10} for 2- to 10-day lead times). After dressing with the new kernel, shown in Fig. 4(b), the rank histogram becomes much flatter throughout 1- to 10-day forecast lead times, which indicates a more reliable ensemble spread. The χ^2 test cannot reject the null hypothesis that the rank histogram is flat even with confidence as low as 88% (the P values greater than 0.12). For the RS-10d-globe dressed ensemble in Fig. 4(c), the rank histogram is over-dispersive through the 1- to 10-day forecast lead times. The χ^2 test confirms this impression of non-uniformity. The P value is nearly zero (much smaller than 10^{-10}) for all lead times, indicating the null hypothesis of uniform rank histogram can be rejected with a high confidence level (much higher than 99%). For the RS03 method, where the best member is identified by RS-1-id-east shown in Fig. 4(d), the histogram is over-dispersive for lead times of 1 to 7 days and the χ^2 test rejects the null hypothesis that the rank histogram is flat with confidence level much higher than 99% (the P value is much smaller than 0.0001). Figure 4(e) is the result corresponding to RS-id-east. The rank histogram is over-dispersive for lead times of 1 to 2 days and under-dispersive for lead times of 8 to 10 days. The χ^2 test confirms the non-uniformity for these five lead times by rejecting the hypothesis of uniformity of rank with a confidence level much higher than 99% (the P value is much smaller than 0.01).

In Fig. 5 we show the Brier score (BS, Brier 1950; Murphy 1973; Wilks 1995) measurement results. Four climatologically equally likely bins are defined by using summer 1999 U_{500} verifications over the 14 verification sites. For each lead time, a probabilistic forecast is made for each of the four bins by the relative frequency of ensemble members falling in each bin. The BS is then calculated (see details in Murphy 1973) using the four bins for each of the 14 sites for each of 92 forecasts of summer 2001 and then averaged over all 14 sites throughout all of the season's forecasts. The number of samples of BS used for averaging for each lead time is thus $14 \times 92 = 1288$. The BS corresponding to using the sample climatology, i.e. the uncertainty term when decomposing the BS, is also shown in Fig. 5. To estimate the significance of the differences between curves, a bootstrap resampling technique (Efron and Tibshirani 1986; Wilks 1995; Hamill 1999; Mullen and Buizza 2001; RS03) is used to estimate the $\pm\sigma$ bounds (i.e. standard error) for each curve. Note that in bootstrap resampling

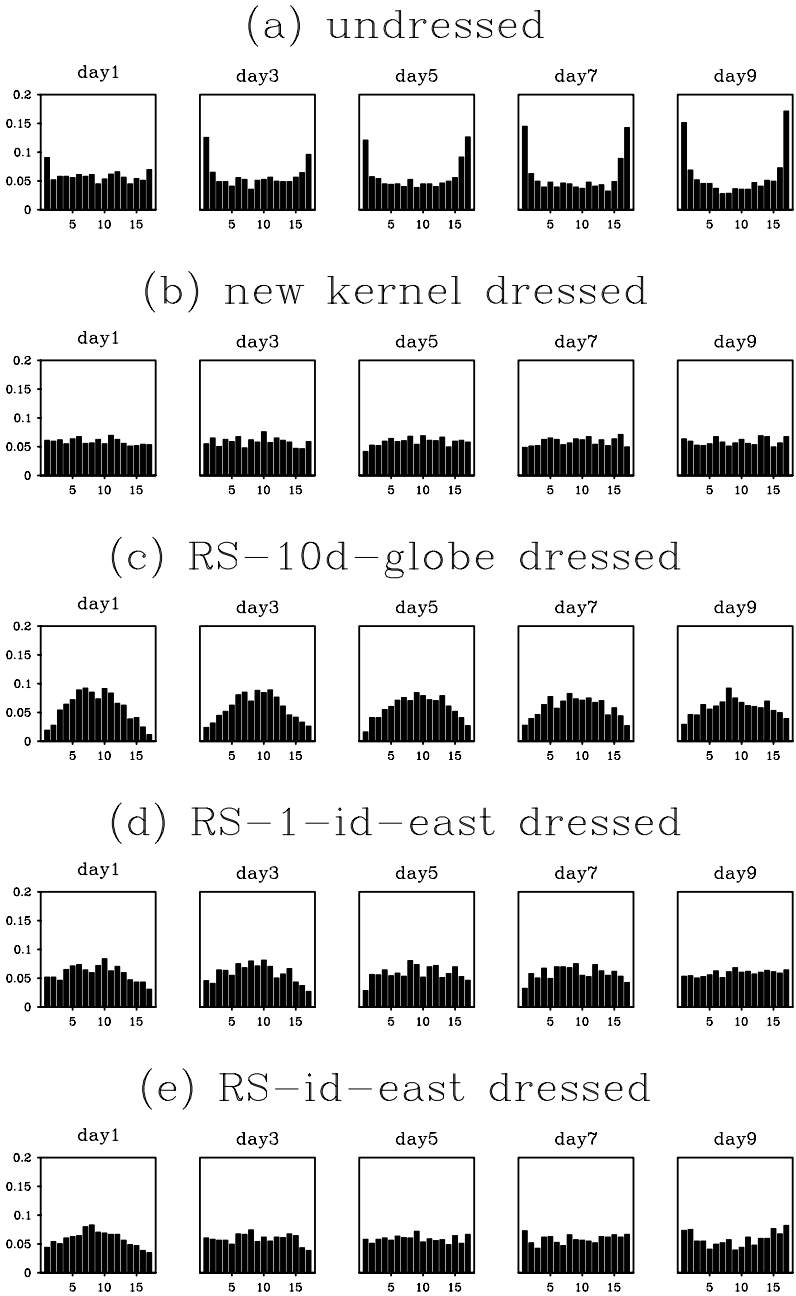

Figure 4. Rank histograms for: (a) undressed, (b) new kernel dressed, (c) RS-10d-globe dressed, (d) RS-1-id-east dressed and (e) RS-id-east dressed CCM3 ETKF 500 hPa zonal wind ensembles over 14 verification sites for lead times of 1, 3, 5, 7 and 9 days. See text for details.

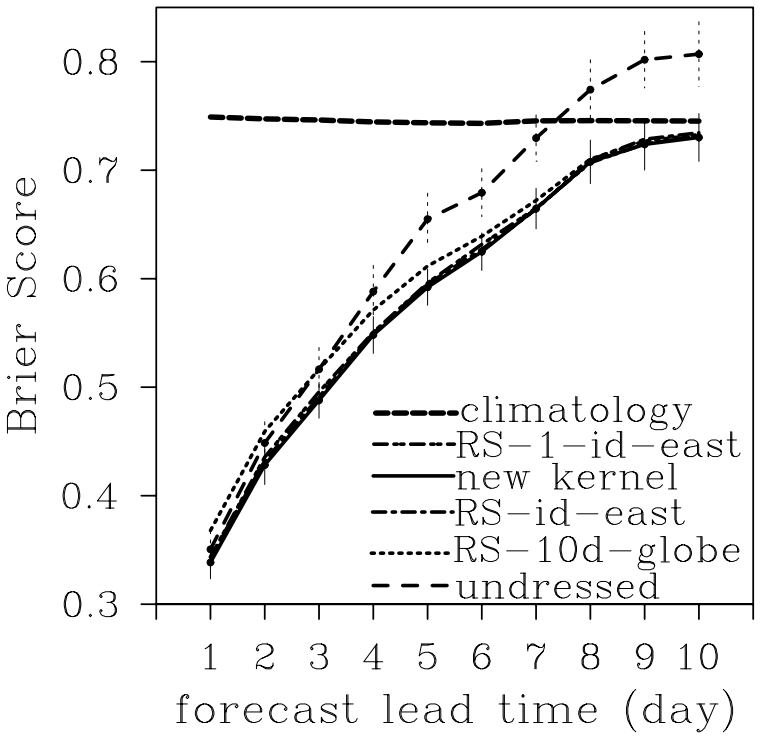

Figure 5. Brier scores (BS) for the undressed, new kernel dressed, RS-10d-globe dressed, RS-1-id-east dressed and RS-id-east dressed CCM3 ETKF 500 hPa zonal wind ensembles from 1- to 10-day lead times. The BS from the sample climatology is also shown. The vertical solid and dashed lines are the standard errors of BS calculations with given samples for the new kernel dressed and undressed ensembles, respectively. See text for further details.

the samples are required to be independent. Since the 1288 BS samples could be spatially and temporally correlated, before resampling we first estimate the number of independent samples within the 1288 BS samples. Following RS03, we divide the total 1288 samples into independent blocks and take the BSs averaged over each block as actual independent samples. We first divide the 14 sites into groups to ensure that the BS time series averaged over each group are uncorrelated among different groups. We end up having three independent groups. Then for each group we work out the length of the temporal block in such a way as to ensure that the autocorrelation of the BS time series given by averaging the BSs over each temporal block is nearly zero. After we obtain the independent samples, 100 bootstrap samples are generated by resampling the independent samples with replacement as recommended by Efron and Tibshirani (1986). These 100 bootstrap samples were used to estimate the $\pm\sigma$ bounds, i.e. the standard error of each curve in Fig. 5.

In Fig. 5, $\pm\sigma$ bounds for the curves of the undressed ensemble and the new dressing kernel are shown. From Fig. 5, the dressed ensemble with new kernel is seen to perform better than the undressed ensemble for 1- to 10-day forecast lead times. The improvements for 4- to 10-day forecasts are statistically significant. It is also better than the best-member dressed ensemble RS-10d-globe for 1- to 10-day lead times with significance for lead times of 1 to 2 days. Decomposition of the BS (Murphy 1973) shows that the improvement of the new method relative to the best-member method is due to the improvement in the reliability component of the BS. The RS-10d-globe

ensemble is worse than the undressed ensemble for lead times of 1 to 2 days. The RS-10d-globe ensemble is significantly better than the undressed ensemble for lead times of 5 to 10 days. The scores for the best-member dressed ensembles, RS-id-east and RS-1-id-east, are statistically indistinguishable from the new kernel dressed ensemble. Note that RS-10d-globe has worse BSs than both RS-id-east and RS-1-id-east, which is inconsistent with the argument from RS03 that full space or high-dimensional space should be used to identify the best member. To explain why the RS-10d-globe ensemble is worse than the RS-id-east and RS-1-id-east ensembles, we first notice that the error variance of the best member defined in RS-10d-globe 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.

We also tried (not shown) using the continuous ranked probability score (CRPS, Hersbach 2000) and the ignorance score (IGN, Roulston and Smith 2002). The comparison results from the CRPS and the IGN are qualitatively the same as that from the BS. Note that in computing these probability scores the ensemble size for the dressed ensemble is 512, which is much larger than the undressed-ensemble size of 16. Thus the improvement of the dressed-ensemble scores relative to the undressed-ensemble scores may partly come from the increase of the ensemble size (Richardson 2001; RS03). This is confirmed when we randomly select 16 out of 512 members to calculate the BS for the dressed ensemble. The results (not shown) show that the improvement of the dressed ensemble relative to the undressed ensemble is smaller than that shown in Fig. 5. The relative performance of the new dressing method versus the best-member method is similar to Fig. 5.

In summary: tests with the CCM3 ETKF ensembles for 500 hPa zonal wind, U_{500}, and other variables at other levels (not shown) show that the performance of the best-member dressed ensemble is highly dependent on the choice of subspace used to define the best member, and that the new dressing kernel can provide a more reliable estimate of the variance of the forecast errors than the best-member dressed ensembles.

5. Application to cooling degree-days forecasts for Boston: A test on forecast error covariance estimates

The rank histogram and BS tests in section 4 only measure the skill of forecasts of individual variables. User specific weather application functions usually depend nonlinearly on more than one weather variable (Palmer 2002). Distributions of such weather application functions are not only sensitive to the forecast error for an individual variable but also sensitive to the covariance of the forecast errors among these variables. As the new kernel augments underdispersive ensembles by way of providing reliable estimates for both the error variance and the error covariance among weather variables of interest, the new kernel is expected to provide reliable ensemble forecasts for such weather application functions. In this section, we demonstrate this property of the new dressing kernel by applying it to the problem of forecasting the accumulative CDDs, a weather index frequently used by users of weather derivatives. Another purpose of this section is to show how ensemble forecasts can be fed into a quantitative user application model, and how the resulting output can be used to form probabilistic forecasts of the economic variable relevant to the user (Palmer 2002).

(*a*) *Cooling degree days definition*

To manage the risks associated with abnormally warm or cool summers, a frequently used weather index is accumulated CDDs (for more information see

http://www.cme.com/prd/wec/abtwthder2766.html of the Chicago Mercantile Exchange). The CDD is defined as:

$$\mathrm{CDD} = \sum_{i=1}^{N_\mathrm{d}} \max(0,\, T_i - 65\,^\circ\mathrm{F}), \tag{15}$$

where N_d is the number of days over which the CDD is accumulated (i.e. the contract period) and T_i is the arithmetic average of the daily maximum and minimum 2 m temperatures in degrees Fahrenheit on the ith day of the period (following Zeng (2000)). Note that knowing the distribution of temperature forecast errors on each of the N_d days defining the CDD is *not* sufficient to determine the probability density function (pdf) of CDDs; one must also know how the temperature errors are correlated through time, because if a temperature error in the day 2 forecast is positively correlated to temperature errors in the day 1 and day 3 forecasts then the distribution of CDDs will be broader than it would be if there were no such correlation.

(*b*) *Application of dressing*

In the following experiment, we only consider samples over a single site, Boston, for one season. In order to increase the number of independent samples, we consider CDDs accumulated over only 3 days. (The Chicago Mercantile Exchange's CDD contracts pertain to CDDs accumulated over a month or a season.). There are two ways to augment the CDD ensemble derived from the 16-member CCM3 ETKF ensemble forecasts for daily 2 m temperature: one is to dress CDD ensemble forecasts directly; the other is to dress T_i and substitute the dressed T_i in Eq. (15). However, if we were to dress CDDs directly we would have to modify our dressing algorithm to account for the fact that CDDs are positive definite. Because of this, and because we want to demonstrate how the new dressing technique can account for correlations of temperature errors through time, we choose to dress T_i. Also, notice that once T_i is dressed it can be applied to other user application functions as well. Specifically, to obtain a dressed-ensemble forecast of the 3-day CDDs, we first dress 1- to 3-day T_i output from the CCM3 ETKF ensemble and then substitute each of the dressed 1- to 3-day T_i forecasts for Boston into Eq. (15).

The CCM3 ETKF T_i outputs are interpolated to the single verification site at Boston. The verifications for CDD and T_i for the summers of 1999 and 2001 are obtained from the Chicago Mercantile Exchange at http://www.cme.com/dta/hist. The training and dressing procedures are similar to those described in subsection 4(a) except: (i) the bias for T_i is computed from the previous 2-weeks' forecasts; (ii) to account for the correlation of errors, the second moment constraint dressing kernel is built by simply placing 1-to-3-day T_i forecasts for Boston and the corresponding verifications in sample vectors with sizes of three elements when constructing the terms in Eq. (9); (iii) the subspace to identify the best member is over Boston from 1- to 3-day lead times and thus the best-member error samples for T_i, $i = 1, 2, 3$ is stored in three-element vectors for archiving the best-member errors; and (iv) the best-member dressing perturbations are drawn from a zero-mean multi-dimensional (three-dimensional in this case) normal distribution whose covariance is consistent with the covariance of the archived best-member errors. With (iv), the number of best-member dressing perturbations drawn is not limited by the length of the time period during which the best-member error archive is built.

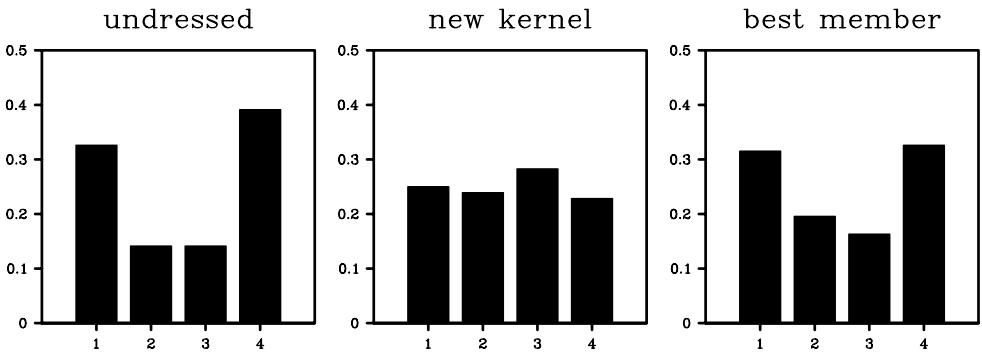

Figure 6. Rank histograms for undressed, new kernel dressed, and the best-member dressed 3-day accumulated cooling degree-day ensembles over Boston during summer 2001.

(*c*) *Results on the reliability of the dressed CDD ensemble spread*

Figure 6 shows the reliability of the spread of the accumulated CDD ensembles measured by the rank histograms. The figure shows that the undressed CDD ensemble underpredicts the CDD forecast uncertainty. After dressing with the best-member method, it is still underdispersive. In comparison, the new dressing kernel can provide reliable spread for the 3-day accumulated CDD forecasts. Note that the number of realizations of verifications for one season's forecasts over a single site is limited for constructing the rank histogram if all ensemble members are used as ranks. To overcome this problem, as in subsection 4(b), we randomly choose a relatively small number of ensemble members out of all the total ensemble members to define the ranks for the rank histogram. The result shown in Fig. 6 corresponds to the case where we randomly choose three members out of all 4096 dressed-ensemble members to build four ranks for each ensemble forecast. Also note, for situations where the verification exactly equals some of the ensemble members, such as CDD forecasts of zero and a verification of zero, the number of members (m) equal to the verification is first counted. Then we assign uniform random numbers between 0 and 1 to the m members and the verification. The m members are ordered according to the assigned random numbers. The rank of the verification is then determined by the rank of the random number assigned to the verification among the m random numbers assigned to the m tied ensemble members. This is similar to the method for constructing rank histograms for precipitation discussed in Hamill and Colucci (1997). The χ^2 test for the uniformity of the rank histogram confirms the flatness of the rank histogram of the new kernel dressed CDD ensembles (P value as large as 0.74) and the non-flatness of those of the undressed (P value as small as 0.0001) and the best-member dressed CDD ensembles (P value as small as 0.02). The underdispersion of the best-member dressed ensemble indicates that the best-member dressing kernel is failing to provide reliable error variance estimates for individual T_i and/or it cannot reliably represent the temporal correlation of forecast errors. We also measure the skills of the CDD ensembles with the IGN. Four climatologically equally likely categories are built from 2001 summer CDD verifications on Boston. The results of the IGN scores for the CDD ensembles are shown in Fig. 7. The smaller the score, the less ignorant of the CDD probabilistic forecast. The statistical t-test (Ott 1993) shows that the IGN for the new kernel CDD ensemble is significantly

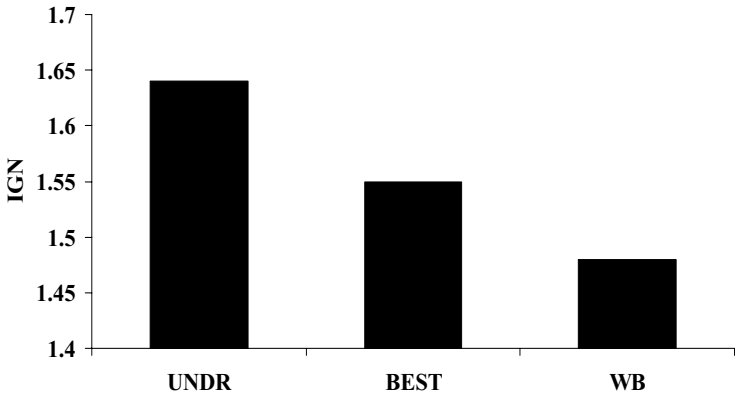

Figure 7. Ignorance scores for the undressed (UNDR), best-member dressed (BEST) and new kernel (WB) dressed cooling degree-day ensembles.

better than those of the best-member CDD ensemble and the undressed CDD ensemble. Therefore, the probabilistic CDD forecast generated from the CDD ensemble augmented by the new dressing kernel is more skilful than the undressed CDD ensemble and the best-member dressed CDD ensemble. The BS measurement (not shown) provides qualitatively the same result. The reliability diagram of the ensemble covariance among the 14 selected sites for U_{500} (not shown) also indicate that the new kernel can provide more reliable forecast error covariance estimates than the best-member method.

6. Conclusion

A new multi-variate dressing method for the purpose of augmenting the spread of underdispersive ensembles has been designed and tested. The method makes distributions from which dressed-ensemble members are drawn indistinguishable from the distribution from which verifying observations are drawn under a seasonally averaged second moment measure. Ensemble bias is first removed, before building training statistics for the dressing kernel and before dressing the current ensembles. The CCM3 ETKF ensemble dressed with the second moment constraint method is more skilful than the corresponding undressed ETKF ensemble. With both a random-number generator experiment and the CCM3 ETKF ensemble framework, the RS03 original best-member dressing method was compared with the second moment constraint dressing method. It was found that the spread of the best-member dressed ensemble can still be underdispersive, or even become overdispersive, depending on such factors as the undressed-ensemble size, how underdispersive the undressed ensemble is and the subspace from which the best member is identified. In contrast, the underdispersive ensembles after they are dressed with the second moment constraint dressing kernel always gave about the right amount of dispersion.

The utility of the second moment constraint dressing relative to the best-member dressing, and the importance of accurately accounting for the temporal correlation of forecast errors, was demonstrated by comparing predictions of accumulative cooling degree-days (CDD) from the ETKF ensemble. It was found that the new second moment constraint dressing kernel provided a 3-day accumulated CDD ensemble with more reliable spread and better skill than the CDD ensemble augmented with the best-member dressing kernel.

In sections 3 and 4 of this paper, the dressing perturbations for the new kernel were drawn from a multivariate normal distribution. As in the best-member method, the dressing perturbations for the new kernel can also be based on an archive of past errors rather than a prescribed distribution. This is achieved by first grouping the historical errors of all ensemble members, and then transforming these errors by pre-multiplying a matrix so as to make the covariance of the transformed errors to be equal to the **Q** matrix in Eqs. (10)–(12). In our experiment, dressing with the archive and the prescribed distribution produce similar results for U_{500} and 2 m temperature. So we only show the results corresponding to the prescribed distribution. Also note that the assumption of a Gaussian dressing kernel is likely to be poor for positive-definite quantities, such as precipitation and 10 m wind speed. To extend the usage of the new dressing kernel for such quantities, a possible option is to transform them in such a way as to make the transformed quantities have more Gaussian type of distributions (Wilks 2002).

In the new dressing method, no dressing is performed for directions where the undressed ensemble is already overdispersive (Eqs. (10)–(12)). In other words, the new method is designed to improve the reliability of underdispersive ensembles, which is one of the common problems in current operational ensembles. Although underdispersion is the most common deficiency of raw ensembles, overdispersion is possible, and any complete ensemble post-processing scheme ought to account for this possibility. To correct overdispersive ensembles, we could try to dress each ensemble member differently. A possible solution would be to dress the central members with more dressing perturbations than the outside members so that the pdf of the dressed ensemble is narrower than the undressed ensemble. We will explore this in future work.

The new dressing kernel, like the best-member kernel, is appropriate for ensembles with each member having similar error statistics. Work is underway to extend the second moment constrained method, discussed here, to the multi-model ensemble case in which differing ensemble members have differing error statistics.

Given large enough datasets, it would be of interest to condition the dressing kernel on flow regimes known to have profound impacts on model error. For example, different dressing kernels might be used on convectively stable and unstable days, and they may also be constructed to be regionally dependent.

ACKNOWLEDGEMENTS

The authors are indebted to the inspiration of Mark Roulston and Leonard Smith who showed us the substantial utility and value of combining statistical and dynamical ensembles. The authors gratefully acknowledge financial supports from ONR grant N00014-00-1-0106, ONR project element 0601153N with project number BE-033-0345, and the Pennsylvania Power and Light. Comments from two reviewers are greatly appreciated.

APPENDIX

(*a*) *Derivation of Eq. (6)*

To derive Eq. (6) first note that using Eq. (4):

$$\begin{aligned}(\boldsymbol{\psi}_{li} - \boldsymbol{\psi}_{lj}) &= (\overline{\mathbf{x}}_l + \mathbf{x}'_{li} + \boldsymbol{\varepsilon}_{li} - \overline{\mathbf{x}}_l - \mathbf{x}'_{lj} - \boldsymbol{\varepsilon}_{lj}) \\ &= ((\mathbf{x}'_{li} - \mathbf{x}'_{lj}) + (\boldsymbol{\varepsilon}_{li} - \boldsymbol{\varepsilon}_{lj})).\end{aligned} \tag{A.1}$$

Using (A.1) on the left-hand side of Eq. (5) gives:

$$\begin{aligned}
&\langle\langle(\boldsymbol{\psi}_{li}-\boldsymbol{\psi}_{lj})(\boldsymbol{\psi}_{li}-\boldsymbol{\psi}_{lj})^{\mathrm{T}}\rangle_{i\neq j}\rangle_l \\
&\quad=\langle\langle((\mathbf{x}'_{li}-\mathbf{x}'_{lj})+(\boldsymbol{\varepsilon}_{li}-\boldsymbol{\varepsilon}_{lj}))((\mathbf{x}'_{li}-\mathbf{x}'_{lj})+(\boldsymbol{\varepsilon}_{li}-\boldsymbol{\varepsilon}_{lj}))^{\mathrm{T}}\rangle_{i\neq j}\rangle_l \\
&\quad=\langle\langle(\mathbf{x}'_{li}-\mathbf{x}'_{lj})(\mathbf{x}'_{li}-\mathbf{x}'_{lj})^{\mathrm{T}}\rangle_{i\neq j}\rangle_l+\langle\langle(\boldsymbol{\varepsilon}_{li}-\boldsymbol{\varepsilon}_{lj})(\boldsymbol{\varepsilon}_{li}-\boldsymbol{\varepsilon}_{lj})^{\mathrm{T}}\rangle_{i\neq j}\rangle_l \\
&\quad=2\langle\boldsymbol{\Sigma}_l^2\rangle_l+2\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle.
\end{aligned}\tag{A.2}$$

Note that the covariance of the dressing perturbations $\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle$ is the same for all ensemble members for all cases; so we put no subscript on this term. Also note that from Eq. (4):

$$(\boldsymbol{\psi}_{li}-\mathbf{y}_l)=(\overline{\mathbf{x}}_l+\mathbf{x}'_{li}+\boldsymbol{\varepsilon}_{li}-\mathbf{y}_l).\tag{A.3}$$

Hence the right-hand side of Eq. (5) is:

$$\begin{aligned}
\langle\langle(\boldsymbol{\psi}_{li}-\mathbf{y}_l)(\boldsymbol{\psi}_{li}-\mathbf{y}_l)^{\mathrm{T}}\rangle_i\rangle_l&=\langle\langle(\overline{\mathbf{x}}_l+\mathbf{x}'_{li}+\boldsymbol{\varepsilon}_{li}-\mathbf{y}_l)(\overline{\mathbf{x}}_l+\mathbf{x}'_{li}+\boldsymbol{\varepsilon}_{li}-\mathbf{y}_l)^{\mathrm{T}}\rangle_i\rangle_l \\
&=\langle\langle((\mathbf{x}'_{li}+\boldsymbol{\varepsilon}_{li})-(\mathbf{y}_l-\overline{\mathbf{x}}_l))((\mathbf{x}'_{li}+\boldsymbol{\varepsilon}_{li})-(\mathbf{y}_l-\overline{\mathbf{x}}_l))^{\mathrm{T}}\rangle_i\rangle_l \\
&=\langle\boldsymbol{\Sigma}_l^2\rangle_l+\langle\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\rangle+\langle(\mathbf{y}_l-\overline{\mathbf{x}}_l)(\mathbf{y}_l-\overline{\mathbf{x}}_l)^{\mathrm{T}}\rangle_l.
\end{aligned}\tag{A.4}$$

Substituting Eqs. (A.1)–(A.4) into Eq. (5) gives Eq. (6).

(*b*) *Derivation of Eq. (9a)*

To derive Eq. (9a), we start with the first term on the right-hand side of Eq. (6). First note that:

$$\begin{aligned}
\langle(\overline{\mathbf{x}}_l^s-\mathbf{y}_l)(\overline{\mathbf{x}}_l^s-\mathbf{y}_l)^{\mathrm{T}}\rangle_l&=\langle((\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)+(\overline{\mathbf{x}}_l-\mathbf{y}_l))((\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)+(\overline{\mathbf{x}}_l-\mathbf{y}_l))^{\mathrm{T}}\rangle_l \\
&=\langle(\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)(\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)^{\mathrm{T}}\rangle_l+\langle(\overline{\mathbf{x}}_l-\mathbf{y}_l)(\overline{\mathbf{x}}_l-\mathbf{y}_l)^{\mathrm{T}}\rangle_l.
\end{aligned}\tag{A.5}$$

Note in deriving the last step in Eq. (A.5), we use the assumption $\langle(\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)(\mathbf{y}_l-\overline{\mathbf{x}}_l)^{\mathrm{T}}\rangle_l=0$, which means the difference between the sample ensemble mean and the underlying ensemble mean does not co-vary with the difference between the verifications (e.g. observations) and the underlying ensemble mean over seasonal forecasts. Also recall that $\langle(\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)(\overline{\mathbf{x}}_l^s-\overline{\mathbf{x}}_l)^{\mathrm{T}}\rangle=\boldsymbol{\Sigma}_l^2/K$. Then from Eq. (A.5), the first term on the right-hand side of Eq. (6) can be approximated as:

$$\langle(\overline{\mathbf{x}}_l-\mathbf{y}_l)(\overline{\mathbf{x}}_l-\mathbf{y}_l)^{\mathrm{T}}\rangle_l=\langle(\overline{\mathbf{x}}_l^s-\mathbf{y}_l)(\overline{\mathbf{x}}_l^s-\mathbf{y}_l)^{\mathrm{T}}\rangle_l-\frac{1}{K}\langle\boldsymbol{\Sigma}_l^2\rangle_l.\tag{A.6}$$

If we approximate $\boldsymbol{\Sigma}_l^2$ in the last term of Eq. (A.6) and the second term on the right-hand side of Eq. (6) with $\boldsymbol{\Sigma}_l^{s^2}$, then we get Eq. (9a).

REFERENCES

Anderson, J. L. 1996 A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9,** 1518–1530

Atger, F. 1999 The skill of ensemble prediction systems. *Mon. Weather Rev.*, **127,** 1941–1953

2003 Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Weather Rev.*, **131,** 1509–1523

2004 Relative impact of model quality and ensemble deficiencies on the performance of ensemble based probabilistic forecasts evaluated through the Brier score. *Nonlinear Proc. Geophys.*, **11,** 399–409

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

Brier, G. W. 1950 Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78,** 1–3

Du, J., Mullen, S. L. and Sanders, F. 1997 Short-range ensemble forecasting of quantitative precipitation. *Mon. Weather Rev.*, **125,** 2427–2459

Eckel, F. A. and Walters, M. K. 1998 Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather Forecasting*, **13,** 1132–1147

Efron, B. and Tibshirani, R. 1986 Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1,** 54–77

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

Hamill, T. M. 1999 Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecasting*, **14,** 155–167

2001 Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.*, **129,** 550–560

Hamill, T. M. and Colucci, S. J. 1997 Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.*, **125,** 1312–1327

1998 Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Weather Rev.*, **126,** 711–724

Hamill, T. M., Whitaker, J. S. and Wei, X. 2004 Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Weather Rev.*, **132,** 1434–1447

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

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

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

Jeffery, T. K., Hack, J. H., Gordon, B. B., Boville, B. A., Briegleb, B. P., Williamson, D. L. and Rasch, P. J. 1996 'Description of the NCAR community climate model (CCM3)'. NCAR Technical Note NCAR/TN-420+STR. National Center for Atmospheric Research, Boulder CO, USA

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, B., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R. and Joseph, D. 1996 The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.*, **77,** 437–471

Kass, R. E. and Raftery, A. E. 1995 Bayes factors. *J. Am. Stat. Soc.*, **90,** 773–795

Krzysztofowicz, R. and Sigrest, A. A. 1999 Calibration of probabilistic quantitative precipitation forecasts. *Weather Forecasting*, **14,** 427–442

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996 The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122,** 73–119

Mullen, S. L. and Buizza, R. 2001 Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Weather Rev*., **129,** 638–663

Murphy, A. H. 1973 A new vector partition of the probability score. *J. Appl. Meteorol.*, **12,** 595–600

Ott, R. L. 1993 *An introduction to statistical methods and data analysis*. Fourth edition. Duxbury Press, Belmont CA, USA

Palmer, T. N. 2002 The economic value of ensemble forecasts as a tool for assessment: From days to decades. *Q. J. R. Meteorol. Soc.*, **128,** 747–774

Raftery, A. E., Balabdaoui, F., Gneiting, T. and Ploakowski, M. 2003 'Using Bayesian model averaging to calibrate forecast ensembles'. Technical report No. 440. Department of Statistics, University of Washington, USA

Richardson, D. S.	2000	Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126,** 649–667
	2001	Measures of skill and values of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.*, **127,** 2473–2489
Roulston, M. S. and Smith, L. A.	2002	Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, **130,** 1653–1660
	2003	Combining dynamical and statistical ensembles. *Tellus*, **55A,** 16–30
Roulston, M. S., Kaplan, D. T., Hardenberg, J. and Smith, L. A.	2003	Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy*, **28,** 585–602
Stensrud, D. J. and Yussouf, N.	2003	Short-range ensemble predictions of 2 m temperature and dew-point temperature over New England. *Mon. Weather Rev.*, **131,** 2510–2524
Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E.	1999	Using ensembles for short-range forecasting. *Mon. Weather Rev.*, **127,** 433–446
Toth, Z. and Kalnay, E.	1993	Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.*, **74,** 2317–2330
	1997	Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.*, **125,** 3297–3319
Toth, Z., Zhu, Y. and Marchok, T.	2001	The use of ensembles to identify forecasts with small and large uncertainty. *Weather Forecasting*, **16,** 463–477
Wang, X. and Bishop, C. H.	2003	A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60,** 1140–1158
Wang, X., Bishop, C. H. and Julier, S. J.	2004	Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Weather Rev.*, **132,** 1590–1605
Whitaker, J. S. and Loughe, A. F.	1998	The relationship between ensemble spread and the ensemble mean skill. *Mon. Weather Rev.*, **126,** 3292–3302
Wilks, D. S.	1995	*Statistical methods in the atmospheric sciences*. Academic Press, San Diego CA, USA
	2002	Smoothing forecast ensembles with fitted probability distributions. *Q. J. R. Meteorol. Soc.*, **128,** 2821–2836
Zeng, L.	2000	Weather derivative and weather insurance: Concept, application, and analysis. *Bull. Am. Meteorol. Soc.*, **81,** 2075–2082
Zhu, Y., Yyengar, G., Toth, Z., Tracton, S. M. and Marchok, T.	1996	'Objective evaluation of the NCEP global ensemble forecasting system'. Pp. J79–J82 in Preprints of the 15th conference on weather analysis and forecasting, Norfolk, VA. American Meteorological Society, Boston, USA
Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K.	2002	The economic value of ensemble-based weather forecasts. *Bull. Am. Meteorol. Soc.*, **83,** 73–83