



A Probabilistic Approach to Adaptive Covariance Localization for Serial Ensemble Square Root Filters

YICUN ZHEN

Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania

FUQING ZHANG

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 9 December 2013, in final form 10 July 2014)

ABSTRACT

This study proposes a variational approach to adaptively determine the optimum radius of influence for ensemble covariance localization when uncorrelated observations are assimilated sequentially. The covariance localization is commonly used by various ensemble Kalman filters to limit the impact of covariance sampling errors when the ensemble size is small relative to the dimension of the state. The probabilistic approach is based on the premise of finding an optimum localization radius that minimizes the distance between the Kalman update using the localized sampling covariance versus using the true covariance, when the sequential ensemble Kalman square root filter method is used. The authors first examine the effectiveness of the proposed method for the cases when the true covariance is known or can be approximated by a sufficiently large ensemble size. Not surprisingly, it is found that the smaller the true covariance distance or the smaller the ensemble, the smaller the localization radius that is needed. The authors further generalize the method to the more usual scenario that the true covariance is unknown but can be represented or estimated probabilistically based on the ensemble sampling covariance. The mathematical formula for this probabilistic and adaptive approach with the use of the Jeffreys prior is derived. Promising results and limitations of this new method are discussed through experiments using the Lorenz-96 system.

1. Introduction

The Kalman filter was first proposed by Kalman (1960) under the assumption of linear dynamics, knowledge of certain covariance information, and Gaussian random variables. Evensen (1994) first introduced the ensemble Kalman filter (EnKF) into this context. EnKF takes advantage of a short-term ensemble to estimate flow-dependent background error covariance, the computational cost of which is much more affordable than the original Kalman filter for large dimensional computation. As variations of EnKF, several ensemble square root filters (EnSRF) have been developed [for a summary see Tippett et al. (2003)] in order to maintain consistency between the ensemble and the covariance

(see Houtekamer and Mitchell 2001; Anderson 2001; Bishop et al. 2001; Whitaker and Hamill 2002; Snyder and Zhang 2003). Ott et al. (2004) designed the local ensemble transform Kalman filter so the state variables can be updated completely in parallel. Meanwhile people are devoting efforts to combining EnKF with three- and four-dimensional variational data assimilation (3D/4D-Var; e.g., Hamill and Snyder 2000; Lorenc 2003; Wang et al. 2008a,b; Zhang et al. 2009, 2013). In the ensemble Kalman filters, sampling error is one of the common problems.

Since the ensemble size is usually far less than the state dimension in geoscience applications, it is impossible for the sample covariance calculated by the ensemble members to capture all the main properties of the true covariance. Some of the sampling errors can be eliminated when the physical correlation radius is small compared to the spatial dimension of the state, which is usually the case in numerical weather prediction. Houtekamer and Mitchell (2001) and Hamill et al.

Corresponding author address: Fuqing Zhang, Department of Meteorology, The Pennsylvania State University, 503 Walker Building, University Park, PA 16802.
E-mail: fzhang@psu.edu

TABLE 1. Algorithm of sequential localized ensemble square root Kalman filter: Algorithm AK.

Input	$\bar{\mathbf{x}}^f \in \mathbb{R}^{n \times 1}$ $\boldsymbol{\varepsilon}^f \in \mathbb{R}^{n \times N}$ $\mathbf{y}^o \in \mathbb{R}$ $\alpha \in [0, 1]$ $\mathbf{R} \in \mathbb{R}$ $\mathbf{h} \in \mathbb{R}^{1 \times n}$ $\mathbf{d} \in \mathbb{R}^{n \times 1}$ $\text{ROI} \in \mathbb{N}$	Mean of ensemble forecast Perturbations of ensemble members Scalar observation Relaxation coefficient Observational variance Observational operator Distance between state variables and the observation Localization radius (used in GC function)
Assimilation	Step 1: Step 2: Step 3: Step 4:	$\mathbf{B}^s = \frac{1}{N-1} \boldsymbol{\varepsilon}^f (\boldsymbol{\varepsilon}^f)^T, \mathbf{B}^s \in \mathbb{R}^{n \times n}$ $\boldsymbol{\rho} = \text{GC}(\text{ROI}, \mathbf{d}), \boldsymbol{\rho} \in \mathbb{R}^{n \times 1}$ $\mathbf{W} = \frac{1}{1 + \sqrt{\frac{\mathbf{R}}{\mathbf{R} + \mathbf{h}\mathbf{B}^s\mathbf{h}^T}}}, \mathbf{r}^s = \frac{\mathbf{B}^s\mathbf{h}^T}{\mathbf{R} + \mathbf{h}\mathbf{B}^s\mathbf{h}^T}, \Delta\mathbf{y} = \mathbf{y}^o - \mathbf{h}\bar{\mathbf{x}}^f$ $\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + (\mathbf{r}^s \cdot \boldsymbol{\rho})\Delta\mathbf{y}, \boldsymbol{\varepsilon}^a = [\mathbf{I} - (\mathbf{r}^s \cdot \boldsymbol{\rho})\mathbf{W}]\boldsymbol{\varepsilon}^f$
Output	$\bar{\mathbf{x}}^a \in \mathbb{R}^{n \times 1}$ $\boldsymbol{\varepsilon}^a \in \mathbb{R}^{n \times N}$	Mean of ensemble analysis Analysis perturbation
After having assimilated all observations, do the step of relaxation:		
	$\boldsymbol{\varepsilon}^a \leftarrow \alpha \boldsymbol{\varepsilon}^f + (1 - \alpha)\boldsymbol{\varepsilon}^a$	

(2001) used localization to remove spurious correlations between distant variables. The localization method can be used in all ensemble-based filters and it has been shown to be a powerful tool for limiting sampling error due to a small ensemble size. The Gaspari–Cohn (GC) localization function (Gaspari and Cohn 1999) is symmetric positive definite and widely used for a wide range of applications. Alternative methods include Jun et al. (2011) who considered estimating the background covariance by smoothing the sample covariance using a kernel function. The GC function is effective for spatial localization, but the localization radius in most applications until recently is chosen empirically. Anderson (2012) sheds some new light on this problem by making use of a probabilistic approach. Anderson and Lei (2013) further extended the probabilistic approach to determining the localization radius based on observation impact. Lei and Anderson (2014) chose the localization radius to be the one that minimizes the difference between sample covariance and localized sample covariance computed from different samples. Meanwhile Bishop and Hodyss (2007) and Bishop et al. (2011) proposed the use of entry by entry powers of sample correlation together with a non-adaptive empirical localization function. The method proposed in the current study is based on a sequential ensemble square root filter. For covariance inflation we utilize the relaxation method from Zhang et al. (2004). To clarify, we present the algorithm in detail in Table 1. Moreover we would like to mention that the method under discussion in this manuscript only applies to the case when observations are uncorrelated. Further

generalizations of this approach to correlated observation errors will be the subject of future research.

The structure of this paper is as follows. In section 2 we propose a variational approach through minimizing a cost function to find the optimum localization radius in the case that the true covariance is known. In section 3 we generalize the method to the case where the true covariance is unknown but can be represented probabilistically from the sample covariance. Numerical results about the performance of this method are presented through sections 2, 3, and 4. In particular, section 4 contains the numerical results based on the Lorenz-96 system. Section 5 provides a summary and discussion on the proposed localization method.

2. Optimal localization with true covariance: A cost function approach

As mentioned earlier, when the physical correlation radius is small the sampling noise at far distances can be largely reduced by localization. Nevertheless, the benefits of this process can occur at the expense of removing small but nonnegligible true covariance beyond the cutoff distance and lowering the true covariance within the localization radius. Figure 1 is a visual illustration of the effect of covariance localization. In this experiment, we use the GC correlation function as the true covariance by setting the true radius of influence (ROI) equal to 5. We draw 61 sample vectors from the Gaussian distribution $\mathcal{N}(0, \mathbf{B}^{\text{true}})$ and compute the sample covariance. Using a localization radius of 24, the localized sampling covariance is then

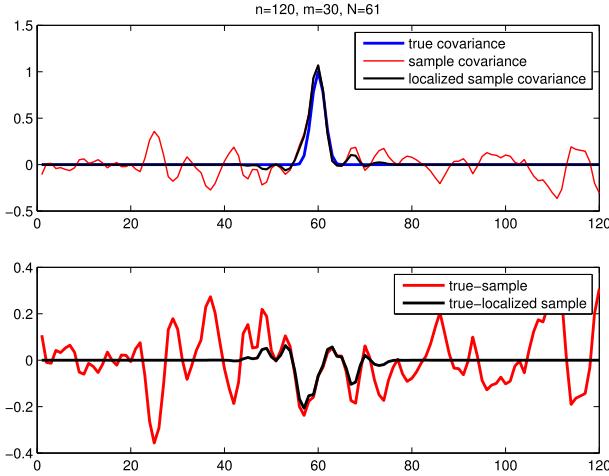


FIG. 1. (top) Gaspari–Cohn covariance (blue line), which serves as the true covariance with radius of influence equal to 5; ensemble covariance (red line) with ensemble size $N = 61$; and localized sample covariance (black line) with localization radius 24. (bottom) The difference between the true covariance and the sample/localized sample covariance. This panel shows that localization can significantly reduce sampling error when the true correlation radius is far smaller than the dimension of the domain.

compared with the raw sampling covariance and the true covariance. We can clearly see that the localization method effectively reduces the spurious correlation at far distances.

Cost function F_0

To gain further understanding of this problem we first examine the solution in the case where the true covariance is known, but we acknowledge that this is impossible in practice. In later sections of this paper we show how to remove this restriction and apply this technique to real problems where the true covariance or Kalman gain is not known. For this scenario we construct a way of describing how “good” the localization radius is for a given ensemble by associating it with the value of a cost function F_0 . Recall, when one observation \mathbf{y} is available, the updating formula of state variable \mathbf{x}_i using the algorithm in section 1 is as follows:

$$\Delta x_i^a = r_i^s \rho_i \Delta \mathbf{y},$$

where Δx_i^a is the mean update of the i th state variable, $\Delta \mathbf{y}$ is the innovation, r_i^s is the regression coefficient calculated from the sample, and ρ_i is the value of localization function for the i th state variable and observation \mathbf{y} .

If we know the true covariance \mathbf{B}^{true} we can simply substitute \mathbf{B}^{true} for \mathbf{B}^s without doing localization. Hence, the exact Kalman updating formula becomes

$$\Delta x_i^a = r_i^t \Delta \mathbf{y}, \quad (1)$$

TABLE 2. Algorithm A0.

Input	
$n, N, m, \text{ROI}_{\max}$	
$r_{ji}^s, r_{ji}^t, d_{ji}$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$	
For $j = 1:m$	
For $\text{ROI} = 1:\text{ROI}_{\max}$	
Compute $F_0(\text{ROI})$ for j th observation;	
End	
Find ROI_j that minimizes $F_0(\text{ROI})$ for the j th observation;	
End	
Collect all ROI_j for $j = 1, 2, \dots, m$ and estimate the probability density of ROI ;	
$\text{ROI}^{(0)}$ = the maximum likelihood estimate of ROI ;	
Output	
$\text{ROI}^{(0)}$	

where

$$r_i^t = \frac{(\mathbf{B}^{\text{true}} \mathbf{h}^T)_i}{\mathbf{R} + \mathbf{h} \mathbf{B}^{\text{true}} \mathbf{h}^T} \quad (2)$$

is the true regression coefficient (perfect Kalman gain). Here \mathbf{h} is the observational operator and \mathbf{R} is the observational covariance.

A natural way of comparing the difference between using the true covariance and the localized sample covariance is represented by the following cost function:

$$F_0(\text{ROI}) = \sum_{i=1}^n (r_i^s \rho_i - r_i^t)^2. \quad (3)$$

The optimal localization radius should be the ROI that minimizes the cost function F_0 . Now let us consider the case when we have more than one observation available at the same time. We can find the optimal ROI for each observation and then find the maximum likelihood estimate of the ROI and choose it to be our overall optimal ROI.

Based on the above analysis, we propose the algorithm A0 (Table 2) (while its results are shown in the left part of Table 3) for evaluating the optimal ROI. In algorithm A0, r_{ji}^s denotes the sample regression coefficient for the j th observation and i th grid point. Similarly r_{ji}^t is the regression coefficient computed by using the true covariance for the j th observation and i th grid point; d_{ji} is the distance (in number of grid points) between the j th observation and i th grid point. In the left half of Table 3 and the left panel of Fig. 2, we show the value of $\text{ROI}^{(0)}$ in terms of grid points for different true covariances and ensemble sizes. Given true covariance of GC form with correlation radius $\text{ROI} = 2, 5, 10, \text{ or } 20$, we draw samples of size $N = 11, 21, \dots, 121$ from the distribution $\mathcal{N}(0, \mathbf{B}^{\text{true}})$ and compute the sample covariance \mathbf{B}^s . We assume observations exist at every grid point and have an error variance of 0.04. Then we compute the $\text{ROI}^{(0)}$ using algorithm A0 for each pair of ROI and N values.

TABLE 3. Comparison of F_0 and F for given true covariance. We do 1000 tests for each experiment and estimate the probability density of $\text{ROI}^{(1)}$ for each experiment. We plot the (rescaled) probability density function of $\text{ROI}^{(1)}$ in the right panels of Fig. 2 and list the maximum likelihood estimates of $\text{ROI}^{(1)}$ in the right half of Table 5. ROI_{\max} is the maximal ROI allowed by algorithm A0 and algorithm A1, respectively. At the first glance we see that for some set of parameters the difference between $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ is large. But we cannot expect that algorithm A0 and algorithm A1 give close values in all situations because in algorithm A1 the information of true covariance is completely unknown anyway. On the other hand, we think the value of $\text{ROI}^{(1)}$ in these cases are still acceptable and sometimes it is close to $\text{ROI}^{(0)}$. Moreover we see that when ensemble size is small $\text{ROI}^{(1)}$ is also small though the true covariance has larger ROI. This is intuitively right since when ensemble size is smaller we put fewer trust on the ensemble covariance. We also see that for larger ensemble size, both algorithms results in larger ROI.

N	$\text{ROI}^{(0)}$ computed by F_0 using true covariance					$\text{ROI}^{(1)}$ computed by F using Jeffreys prior				
	11	21	31	61	121	11	21	31	61	121
ROI = 2	2.5	3.1	3.5	4.2	5	2.2	3.5	4.6	7	17
ROI = 5	7	8.2	8.8	10.4	11.9	3.5	6	7.9	12.3	21.5
ROI = 10	13.2	15.6	17.3	19.8	22.9	4	7.8	10.9	17.7	28.5
ROI = 20	26.5	30.7	33.8	39	45.8	4	9	13	23.5	39.5
ROI_{\max}	130	130	130	130	130	4	9	14	29	59

We do this test 1000 times for each pair of parameters and estimate the probability density of $\text{ROI}^{(0)}$ and plot the curves of (rescaled) probability density in Fig. 2. In Table 3 (left half) we show the maximum likelihood estimate of $\text{ROI}^{(0)}$ for each choice of ensemble size and radius of influence of the true covariance.

For all choices of parameters in this experiment, the resulting maximum likelihood estimate of $\text{ROI}^{(0)}$ is always larger than the true ROI. This is desirable since our goal is to find a localization that can reduce the spurious correlation at a far distance while also approximating the true correlation at a near distance. Second, we can see that for a larger ensemble size (or ROI) the resulting $\text{ROI}^{(0)}$ also gets larger. We also find that as the true ROI increases for a fixed ensemble size, the increment of $\text{ROI}^{(0)}$ is almost linearly proportional to the increment of ROI. That is likely due to the Gaussian assumption. More discussions on Table 3 and Fig. 2 will be presented in section 3.

3. A probabilistic method of localization

In more common and realistic cases the true covariance is almost always unknown so the algorithm A0 is quite limited for real-data implementations. Anderson (2012) introduced a probabilistic method by assuming that the true correlation between a scalar variable and a scalar observation is a random variable based on the ensemble. And in Lei and Anderson (2014) they choose the localization value that minimizes the difference between the sample regression coefficient and the localized sample regression coefficient from another pair of the same type of variables from another ensemble group of the hierarchical ensemble group filter. We follow a similar approach to Anderson (2012) in this study. More precisely speaking, we compute the optimal localization radius whenever any observation is

available for data assimilation. We assume that the true covariance between a scalar observation and all nearby state variables follows a specified error distribution so we can compute the average value of some cost function $F(r_i^t, r_i^s, d_i, \text{ROI})$ over that distribution. Part of the cost function we will present has some resemblance to Eq. (1) in Lei and Anderson (2014). However, the main technique we employ in this manuscript is considerably different from that in Lei and Anderson (2014), which does not assume any prior distribution of the true covariance but instead compares the difference between the sample regression coefficient with the localized sample regression coefficient of the same type, under their definition, of pairs of observation variables and state variables from a different ensemble group. We use the Jeffreys prior and the Wishart distribution to calculate the conditional distribution of the true covariance matrix rather than using the sample covariance computed from another ensemble group member as in Lei and Anderson (2014). It may also be worth noting that the method in Lei and Anderson (2014) finds the spatially homogeneous localization while the scheme we propose in this manuscript can be applied for any location that is spatially varying. In our proposed method of this section $r_i^t = r_i^t(\mathbf{B})$ is still the true regression coefficient, but it is now a function of the true covariance \mathbf{B} represented by a random matrix. Therefore, the regression coefficient r_i^t is a random number. The distribution of the true covariance for a given sample comes directly from Bayes's theorem:

$$p(\mathbf{B} | \mathbf{B}^s) = \frac{p(\mathbf{B}, \mathbf{B}^s)}{p(\mathbf{B}^s)} = \frac{p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B})}{p(\mathbf{B}^s)}. \quad (4)$$

In the above formula $p(\mathbf{B}^s)$ is a constant for a given sample and $p(\mathbf{B}^s | \mathbf{B})$ is the well-known Wishart distribution. Since we assume the prior distribution is

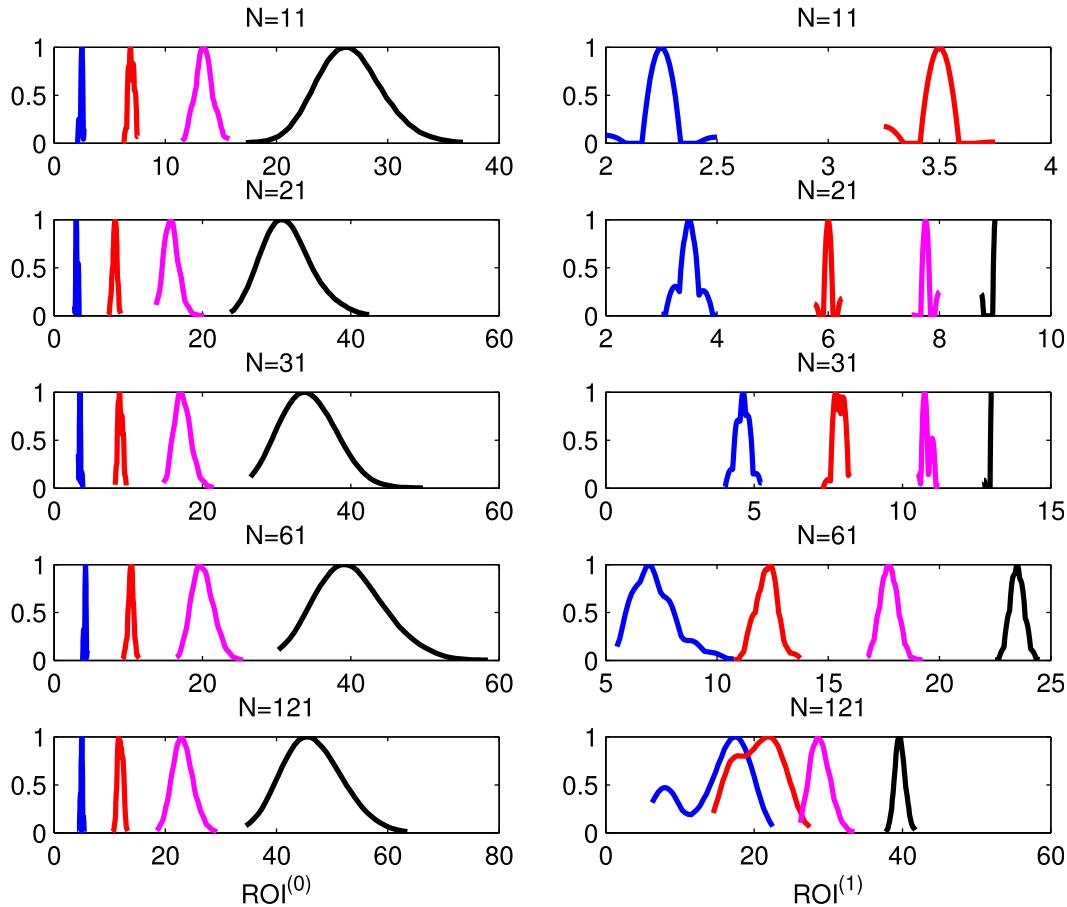


FIG. 2. The (rescaled) probability density estimate of $ROI^{(0)}$ and $ROI^{(1)}$ for experiments using known true covariance with $ROI = 2, 5, 10$, or 20 . Here we normalize the curves so the maximum of each curve is always equal to 1. (left) $ROI^{(0)}$ and (right) $ROI^{(1)}$. Blue curves are for $ROI = 2$. Red curves are for $ROI = 5$. Magenta curves are for $ROI = 10$. Black curves are for $ROI = 20$. The maximum likelihood estimate of each curve is listed in Table 3. The MLE of the curves in (left) is listed in the left half of Table 3. Note that there is no black and magenta curve in the (top right) panel; please refer to the text for a discussion of this.

Gaussian, it has been mathematically proved that the distribution of sample covariance is exactly the Wishart distribution (see Muirhead 1982); $p(\mathbf{B})$ is called the prior distribution, the choice of which is not unique. Here we use the Jeffreys prior, which is proportional to the square root of the determinant of the Fisher information. We do not claim that Jeffreys prior is the most appropriate distribution, but this is commonly used in the statistics community (see Bernardo and Smith 1994).

The average of the cost function F over all true (random) covariance is

$$\int \sum_{i=1}^n F(r_i^t, r_i^s, d_i, ROI) \frac{p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B})}{p(\mathbf{B}^s)} d\mathbf{B}. \quad (5)$$

One limitation is that the Wishart distribution is valid only when the sample size is larger than the dimension of the state variable, which cannot be satisfied in most of the

applications. If the sample covariance is not of full rank, then the space of \mathbf{B}^s is not an open subset of the space of all symmetric positive definite (SPD) matrices. In this case, the space of singular semipositive definite matrix is a surface of the full space of SPD matrix. And the probability distribution lies on this surface. But the true covariance is usually of full rank, hence, it does not lie in this subspace. Though people have studied the analog of the Wishart distribution in the case that \mathbf{B}^s is of deficient rank, we do not use it for the following two reasons:

- 1) the true covariance may not lie in this subspace and the subspace is not flat;
- 2) mathematically it is harder to simplify the cost function to a computable form.

Therefore, we consider computing the truncated average of the cost function instead of the overall average. Specifically speaking, we consider

$$\int \sum_{d_i < \text{ROI}} F(r_i^t, r_i^s, d_i, \text{ROI}) \frac{p(\mathbf{B}_{\text{Loc}}^s | \mathbf{B}_{\text{Loc}}) p(\mathbf{B}_{\text{Loc}})}{p(\mathbf{B}_{\text{Loc}}^s)} d\mathbf{B}_{\text{Loc}}, \tag{6}$$

where $\mathbf{B}_{\text{Loc}}^s$ (or \mathbf{B}_{Loc}) is the sample (or true) covariance of the state variables that are within distance ROI from the observation. ROI can only be chosen among those that the amount of state variables within the distance ROI from an observation is strictly less than the ensemble size N . Intuitively this truncated average only takes into account the influence of the observation on the state variables that lie in the “radius of influence.” This restriction also causes the resulting ROI to be extremely small if we directly apply this method to 3D GCMs. This aspect will be studied further in the future.

a. How to choose the cost function

If we use exactly the same cost function F_0 for the truncated average as in section 2, where the true covariance is known, we find that the variational method often leads to a trivial solution (ROI = 0, not shown). We believe part of the reason is that the truncation ignores all information outside of the radius of influence. In other words, the localization radius resulted by using the function F_0 in Eqs. (5) and (6) might be much different if we instead use a truncated version of F_0 . To limit the loss of information, we consider a modified cost function:

$$F_{11}(r_i^s, r_i^t, d_i, \text{ROI}) = \sum_{d_i < \text{ROI}} [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2]. \tag{7}$$

The advantages of the function F_{11} are the following:

- 1) The truncated F_{11} and the full F_{11} have the same value:

$$\sum_{i=1}^n [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2] = \sum_{d_i < \text{ROI}} [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2]. \tag{8}$$

- 2) In the case the true covariance is known, F_0 and F_{11} only differ by a constant. More precisely speaking,

$$\sum_{i=1}^n [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2] = C + \sum_{i=1}^n (\rho_i r_i^s - r_i^t)^2, \tag{9}$$

where this constant C has no impact on where the cost function’s minimum is taken in this case. Therefore, the constant C has no impact on determining the localization radius.

Combining the above properties we can see that for a given true covariance, substituting F_{11} for F_0 in

TABLE 4. Algorithm A1.

Assumption:	Each row of observational operator is of the form (0, 0, . . . , 0, 1, 0, . . . , 0)
Input	$\mathbf{g}^f \in \mathbb{R}^{n \times N}$ $n \in \mathbb{N}$ $N \in \mathbb{N}$ $\mathbf{R} \in \mathbb{R}$ $(i_1, \dots, i_m) \in \mathbb{N}^m$ $\mathbf{d} \in \mathbb{R}^{n \times 1}$ For $j = 1: m$ ROI _{max} = min[floor($n/2$) + 1, ($N - 3$)/2]; For ROI = 1: ROI _{max} Compute n_{Loc} ; Compute ρ_i for $d_i < \text{ROI}$; Compute r_i^f for $d_i < \text{ROI}$; Compute \mathbb{E}_1 and \mathbb{E}_2 ; Compute $F_1(\text{ROI})$ using Eq. (4); Compute $F_2(\text{ROI})$ using Eq. (5); Compute $F(\text{ROI}) = F_1(\text{ROI}) + F_2(\text{ROI})$; End Find ROI _{j} that minimizes function F ; End Collect all of ROI _{j} and estimate the probability density function for ROI; Find the maximum likelihood estimate ROI ⁽¹⁾ ; ROI ⁽¹⁾
Output	ROI ⁽¹⁾

algorithm A0 is mathematically equivalent. More importantly, in the case that the true covariance is unknown, the truncated average when use Jeffreys prior can be simplified explicitly so the integration of thousands of variables can be avoided. However, because of the use of the uninformative Jeffreys prior as the true covariance probability distribution function, we find that an additional penalty is necessary to be included in the cost function to avoid a trivial solution where the resulting ROI is simply the largest possible radius of influence allowed by the algorithm. Therefore, we define

$$F_{10}(r_i^s, r_i^t, d_i, \text{ROI}) = \sum_{d_i < \text{ROI}} [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2] + \sum_{d_i < \text{ROI}} \rho_i^2 (r_i^s - r_i^t)^2, \tag{10}$$

where the second term is a penalty term that can be interpreted as the sampling noise or stability term that accounts for the difference between the true covariance and sampling covariance scaled by the localization coefficient.

Now we present the complete form of the cost function that we designed for the case when true covariance is unknown:

$$F(\text{ROI}) = \int \left\{ \sum_{d_i < \text{ROI}} [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2] + \sum_{d_i < \text{ROI}} \rho_i^2 (r_i^s - r_i^t)^2 \right\} \frac{p(\mathbf{B}_{\text{Loc}}^s | \mathbf{B}_{\text{Loc}}) p(\mathbf{B}_{\text{Loc}})}{p(\mathbf{B}_{\text{Loc}}^s)} d\mathbf{B}_{\text{Loc}}, \tag{11}$$

where $\rho_i = \text{GC}(\text{ROI}, d_i)$.

For the exact mathematical expression/definition of $p(\mathbf{B}_{\text{Loc}}^s | \mathbf{B}_{\text{Loc}})$, $p(\mathbf{B}_{\text{Loc}})$, $p(\mathbf{B}_{\text{Loc}}^s)$, and $d\mathbf{B}_{\text{Loc}}$, please refer to the [appendix](#). Note that this integral is taken over thousands of variables so it cannot be directly computed numerically. Mathematical simplifications to Eq. (11) can be made under certain assumptions. A computable form of this expression is derived in the [appendix](#).

THEOREM 3.1

Suppose there is a scalar observation available at this time and the observational variance is \mathbf{R} . The observational operator $\mathbf{h} = (0, 0, \dots, 0, 1, 0, \dots, 0)$. Let k be the state variable that the observation is taken for. Let n_{Loc} be the number of state variables that are within distance ROI from the observation. Let

$\boldsymbol{\epsilon}_i \in \mathbb{R}^{1 \times N}$ be the forecast perturbation of the i th component of the state variable.

Let

$$\Delta_{ii} = \frac{|\boldsymbol{\epsilon}_i|^2}{|\boldsymbol{\epsilon}_k|^2} - \frac{\langle \boldsymbol{\epsilon}_k, \boldsymbol{\epsilon}_i \rangle}{|\boldsymbol{\epsilon}_k|^2}.$$

Let

$$\mathbb{E}_1 = \mathbb{E}_{\Gamma[(N-n_{\text{Loc}})/2, 1]} \left\{ \frac{(N-1)\mathbf{R} + |\boldsymbol{\epsilon}_k|^2}{2\mathbf{R}x + |\boldsymbol{\epsilon}_k|^2} \right\},$$

$$\mathbb{E}_2 = \mathbb{E}_{\Gamma[(N-n_{\text{Loc}})/2, 1]} \left\{ \frac{(N-1)\mathbf{R} + |\boldsymbol{\epsilon}_k|^2}{2\mathbf{R}x + |\boldsymbol{\epsilon}_k|^2} \right\}^2,$$

$$F_1(\text{ROI}) = \int \sum_{d_i < \text{ROI}} [(\rho_i r_i^s - r_i^t)^2 - (r_i^t)^2] \frac{p(\mathbf{B}_{\text{Loc}}^s | \mathbf{B}_{\text{Loc}}) p(\mathbf{B}_{\text{Loc}})}{p(\mathbf{B}_{\text{Loc}}^s)} d\mathbf{B}_{\text{Loc}}, \quad \text{and}$$

$$F_2(\text{ROI}) = \int \sum_{d_i < \text{ROI}} \rho_i^2 (r_i^s - r_i^t)^2 \frac{p(\mathbf{B}_{\text{Loc}}^s | \mathbf{B}_{\text{Loc}}) p(\mathbf{B}_{\text{Loc}})}{p(\mathbf{B}_{\text{Loc}}^s)} d\mathbf{B}_{\text{Loc}}.$$

Then

$$F_1(\text{ROI}) = \sum_{d_i < \text{ROI}} \{(r_i^s)^2 (\rho_i^2 - 2\rho_i \mathbb{E}_1)\} \quad \text{and} \tag{12}$$

$$F_2(\text{ROI}) = \sum_{d_i < \text{ROI}} \rho_i^2 \left\{ (r_i^s)^2 (1 - 2\mathbb{E}_1 + \mathbb{E}_2) + \frac{\Delta_{ii} |\boldsymbol{\epsilon}_k|^2}{(N - n_{\text{Loc}} - 1) [(N - 1)\mathbf{R} + |\boldsymbol{\epsilon}_k|^2]} \mathbb{E}_2 \right\}, \tag{13}$$

where $\mathbb{E}_{\Gamma[(N-n_{\text{Loc}})/2, 1]} \{f(x)\}$ denotes the expectation of $f(x)$ when x follows the distribution $\Gamma(N - n_{\text{Loc}}/2, 1)$.

Based on the above theorem we propose algorithm A1 ([Table 4](#)) to adaptively determine the localization radius. The computational cost of this algorithm is $O(N^2 m)$.

b. Comparison of $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ for a given true covariance

We present the experimental results in [Table 3](#) and [Fig. 2](#) for a comparison of ROI using the algorithm A0 presented in [section 2](#) [$\text{ROI}^{(0)}$] versus the new probabilistic algorithm A1 [$\text{ROI}^{(1)}$]. The experiment for $\text{ROI}^{(0)}$ is still the same as that presented in [section 2](#), where the true covariance is known. In addition, we add columns 7–11, which are values of $\text{ROI}^{(1)}$ computed by algorithm A1 for the same true covariance and sample covariance that are

used in columns 2–6, though the information of true covariance is completely unknown in algorithm A1. Again we do 1000 tests for each experiment and estimate the probability density of $\text{ROI}^{(1)}$ for each experiment. We plot the (rescaled) probability density function of $\text{ROI}^{(1)}$ on the right panels of [Fig. 2](#) and list the maximum likelihood estimates (MLE) of $\text{ROI}^{(1)}$ in the right panel of [Table 3](#). Note that the last row of [Table 3](#) is the maximum value of $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ allowed by algorithm A0 and A1 respectively. The upper bound for $\text{ROI}^{(0)}$ can be modified arbitrarily with the larger upper bound translating into a larger range of possible solutions. $\text{ROI}_{\text{max}} = 130$ is found to be a large enough searching range for our application. The upper bound for $\text{ROI}^{(1)}$ is not as flexible as in the $\text{ROI}^{(0)}$ case. In this experiment and the experiments in the next section, the maximum of $\text{ROI}^{(1)}$ is chosen to

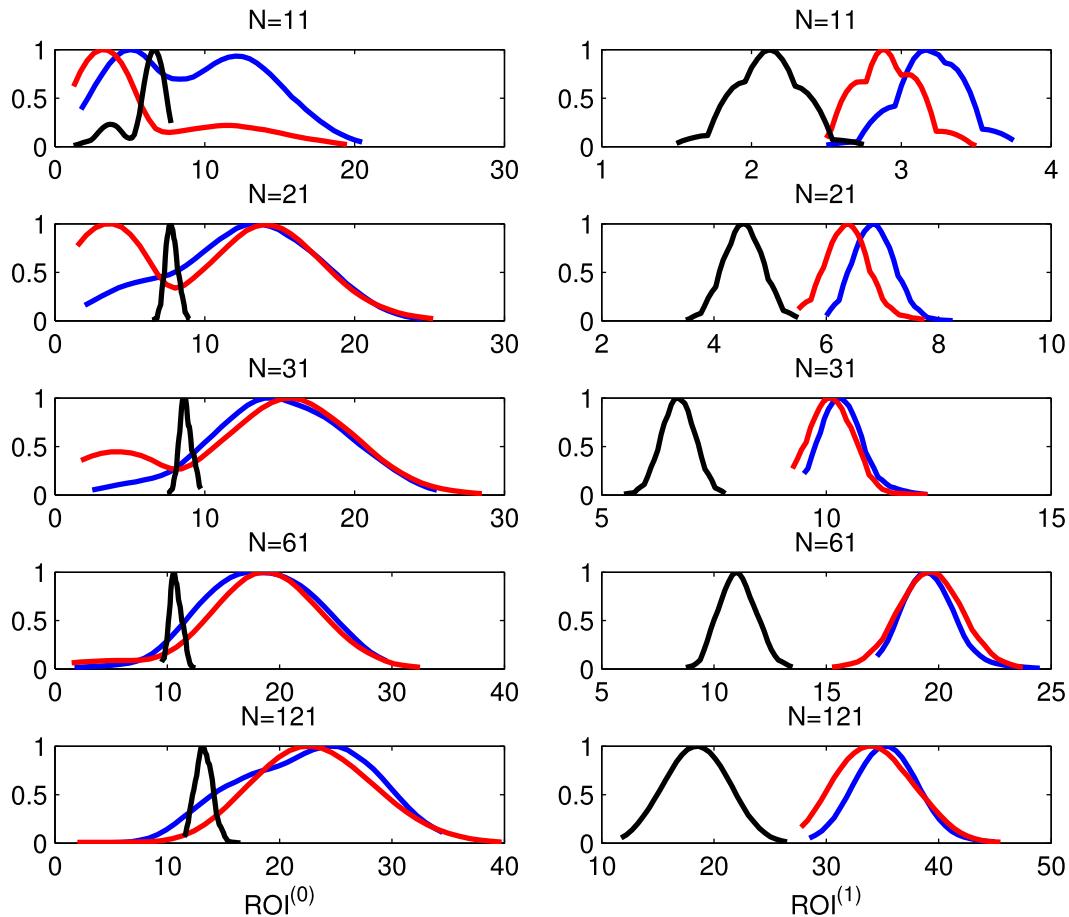


FIG. 3. The (rescaled) probability density estimate of $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ for experiments using approximated true covariance in a Lorenz-96 system with $T_1 = 50$, $T_2 = 1, 10$, and 100 . Here we normalize the curves so the maximum of each curve is always equal to 1. (left) $\text{ROI}^{(0)}$ and (right) $\text{ROI}^{(1)}$. Blue curves are for $T_2 = 1$. Red curves are for $T_2 = 10$. Black curves are for $T_2 = 100$. The MLE of each curve is listed in Table 5.

be $(N - 3)/2$. Therefore, when the ensemble size is small and if the $\text{ROI}^{(1)}$ tends to be large, the $\text{ROI}^{(1)}$ of the 1000 independent tests would simply concentrate at the maximum value allowed by the algorithm. This is why we see some peculiar shape of the black curves in the right panels of Fig. 2 when the ensemble size is small.

When the true $\text{ROI} = 2$, $\text{ROI}^{(0)}$ varies from 2.5 to 5. For the algorithm that uses no knowledge of the true covariance, $\text{ROI}^{(1)}$ changes from 2.2 to 17, which has a much larger range than $\text{ROI}^{(0)}$. For example, when $N = 11$, the MLE of $\text{ROI}^{(0)}$ is about 1.3 times the true ROI . When $N = 121$, the MLE of $\text{ROI}^{(0)}$ is about 2.6 times the true ROI . Note that now there is no similar linear relationship between the MLE and ROI for $\text{ROI}^{(1)}$ [which is found for $\text{ROI}^{(0)}$]. When $N = 11$, $\text{ROI} = 10$ and 20 , the probability density of $\text{ROI}^{(1)}$ all concentrate at a single point, which causes the graph of the (rescaled) probability density function to not be

seen in Fig. 2. Similarly for $N = 21$ and 31 , the probability density of $\text{ROI}^{(1)}$ in Fig. 2 is very concentrated near the maximal possible ROI (i.e., ROI_{\max}). This is because algorithm A1 does not allow $\text{ROI}^{(1)}$ to be larger than $(N - 3)/2$ and the best radius of influence in the search range is very likely to be the upper bound ROI_{\max} . When $N = 121$, the distribution of $\text{ROI}^{(1)}$ is far away from the upper bound ROI_{\max} but the MLE of $\text{ROI}^{(1)}$ does not depend linearly on the true ROI neither.

At first glance, the differences between $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ are quite large over the set of parameters that we examine. When $N = 121$, the MLE of $\text{ROI}^{(1)}$ is larger than that of $\text{ROI}^{(0)}$ when the true $\text{ROI} = 2, 5$, and 10 but smaller than that of $\text{ROI}^{(0)}$ when the true $\text{ROI} = 20$. This result is not surprising because in algorithm A1 the true covariance is completely unknown as in real-data applications. On the other hand, the values of $\text{ROI}^{(1)}$ are close to $\text{ROI}^{(0)}$ when $N = 61$.

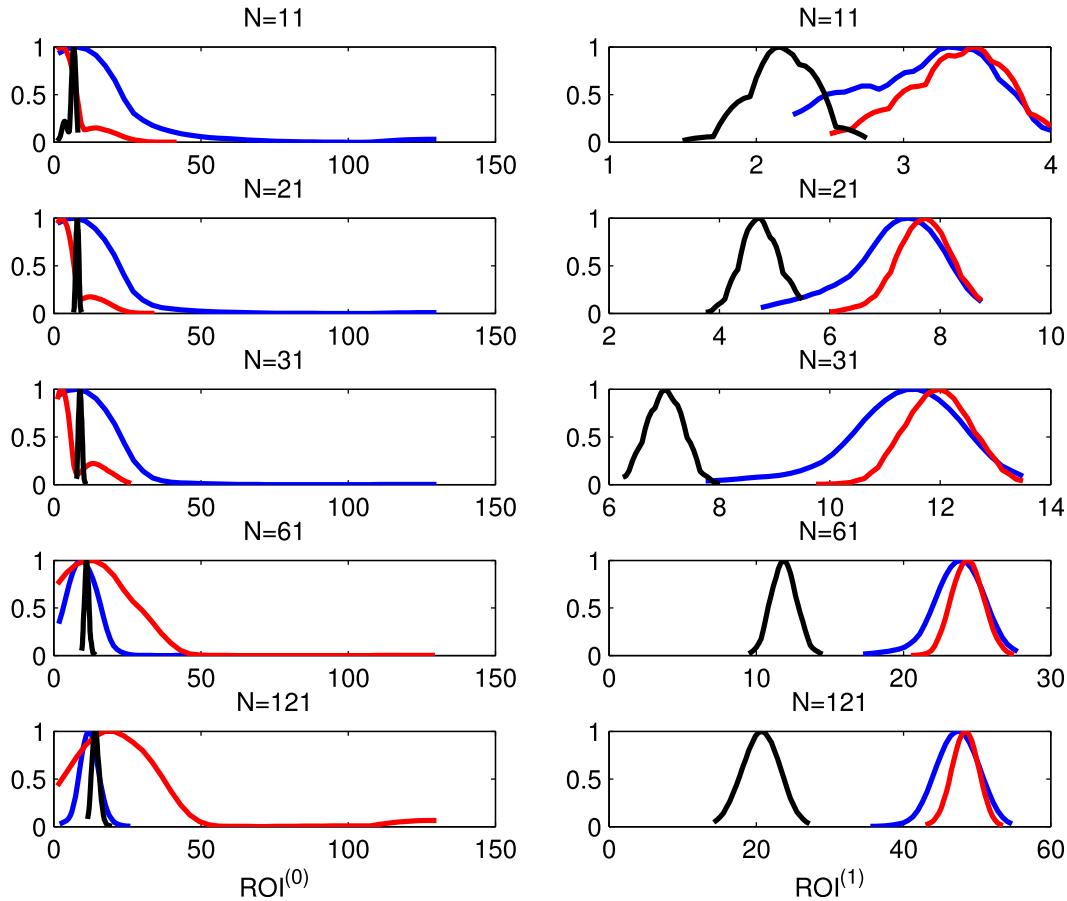


FIG. 4. The (rescaled) probability density estimate of $ROI^{(0)}$ and $ROI^{(1)}$ for experiments using approximated true covariance in a Lorenz-96 system with $T_1 = 500$, $T_2 = 1, 10$, and 100 . Here we normalize the curves so the maximum of each curve is always equal to 1. (left) $ROI^{(0)}$ and (right) $ROI^{(1)}$. Blue curves are for $T_2 = 1$. Red curves are for $T_2 = 10$. Black curves are for $T_2 = 100$. The MLE of each curve is listed in Table 5.

Moreover we see that when the ensemble size is small $ROI^{(1)}$ is also small though the true covariance has larger ROI. We also see that for a larger ensemble size, both algorithms result in larger ROIs. Therefore, the algorithm adjusts the ROI according to the expected sampling error for a given data assimilation cycle.

Although F_{11} would give the same value if we slightly enlarge the region allowed by the Wishart distribution, the integral would be different due to the use of the noninformative Jeffreys prior. As a consequence, if we only use F_{11} as our cost function the resulting $ROI^{(1)}$ would just be ROI_{max} in most situations. This is why the second term in F_{10} is required for making the solution nontrivial. On the other hand, the choice of prior does not necessarily have to be the Jeffreys prior. The penalty term can have other choices as well. In this sense, the method is still quite empirical when the true covariance is unknown. Future studies will explore the use of more informative prior

probability distribution functions for optimizing the ensemble localization distance.

4. Application to the ensemble covariance generated by the Lorenz-96 model

a. Experiments with the Lorenz-96 system and approximated true covariance

Now we design experiments to compare algorithms A0 and A1 in a Lorenz-96 system (Lorenz 2006). Here the system is configured to have 120 variables and 30 uniformly distributed observations, which lie on the model grid points. The external forcing F is set to 8, the time step dt is set to 0.05, and observations appear every two time steps. We use an error variance of 0.04 for all observations in these experiments. Observational operator \mathbf{H} is simply the restriction operator as required by the current version of the algorithm. To get an approximated true covariance in the Lorenz-96 system, we first run 6000 ensemble members for T_1 time steps with the EnKF data assimilation. We then

TABLE 5. Comparison of F_0 and F for approximated, true covariance in Lorenz-96 system. These are maximum likelihood estimates of $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$. The probability estimates are based on 1000 independent tests. Our true covariance \mathbf{B}^{true} is approximated by the sample covariance of 6000 ensemble members. First we run these 6000 members for T_1 time steps with data assimilation. Then we run these members for T_2 time steps without data assimilation. Then we compute the sample covariance and set it to be the true covariance \mathbf{B}^{true} . For each test our sample covariance \mathbf{B}^s is computed from N randomly drawn ensemble members. Then we compute $\text{ROI}^{(0)}$ and $\text{ROI}^{(1)}$ based on that specific sample covariance and true covariance. The left half are maximum likelihood estimates of $\text{ROI}^{(0)}$. The right half are maximum likelihood estimates of $\text{ROI}^{(1)}$. ROI_{max} is the maximal ROI allowed by algorithm A0 and algorithm A1.

N	ROI ⁽⁰⁾ computed by F_0 using approximate true covariance					ROI ⁽¹⁾ computed by F using Jeffreys prior				
	11	21	31	61	121	11	21	31	61	121
$T_1 = 50, T_2 = 1$	5	13.2	14.6	17.7	24.5	3.2	6.8	10.3	19.4	35.1
$T_1 = 50, T_2 = 10$	3.1	3.9	15.8	18.6	22.8	3	6.4	10.2	19.5	33.5
$T_1 = 50, T_2 = 100$	6.6	7.8	8.7	10.6	13.1	2.1	4.5	6.7	11	18.4
$T_1 = 500, T_2 = 1$	7.9	7.7	7.5	9.9	12.2	3.4	7.5	11.5	23.8	47.5
$T_1 = 500, T_2 = 10$	2.3	2.9	2.5	10.9	20.1	3.4	7.7	12	24.4	48.5
$T_1 = 500, T_2 = 100$	6.9	7.9	9	11.2	14.2	2.2	4.7	7.1	11.9	20.8
ROI _{max}	130	130	130	130	130	4	9	14	29	59

run the ensemble members for T_2 time steps without data assimilation. Then we use the whole 6000 ensemble members to get an approximation of the true covariance \mathbf{B}^{true} . In the first T_1 time steps, we use covariance relaxation with $\alpha = 0.5$ (Zhang et al. 2004). Then, we draw N ensemble members from the 6000 members and compute the sample covariance \mathbf{B}^s . Finally we compute the optimal ROI⁽⁰⁾ using algorithm A0 with the known (approximated) true covariance, and compare it with the optimal ROI⁽¹⁾ that was computed by algorithm A1 with no knowledge of the true covariance. We choose $\alpha = 0.5$ here because the model still has a chance to blow up when no localization is used, even if when the ensemble size is 6000. We did 1000 tests for each set of the parameters and we estimated the probability density of the optimum ROI with respect to both algorithm A0 and A1. The graphs of probability density function are shown in Figs. 3 and 4 and the maximum likelihood ROI⁽⁰⁾ and ROI⁽¹⁾ are listed in Table 5.

We do not expect that algorithms A1 and A0 give similar values in all cases, because in algorithm A1 we do not use the true covariance while in algorithm A0 we have the complete information of true covariance.

When $T_1 = 50$ (T_1 is the EnKF analysis duration), $N = 61$, the value of ROI⁽⁰⁾ = 17.7, 18.6, and 10.6 for different T_2 (T_2 is the length of free ensemble forecast without assimilation). We see ROI⁽¹⁾ = 19.4, 19.5, and 11 for different T_2 , respectively, which are very close to the values of ROI⁽⁰⁾. But for $T_1 = 500$ (Fig. 4), we see the probability density function of ROI⁽⁰⁾ is quite peculiar for $T_2 = 1$ and 10 (blue and red curves in the left panels) when the ensemble size is small. The maximum likelihood of ROI⁽⁰⁾ are around 7.7, 2.5, and 8 for different T_2 no matter if $N = 11, 21, \text{ or } 31$. Although it is not clear why $T_2 = 10$ results in ROI⁽⁰⁾ ≈ 2.5 , which might be due to the system dynamics, this shows that ROI⁽⁰⁾ only changes a little when ensemble size changes among

$N = 11, 21, \text{ and } 31$. The estimated probability densities of ROI⁽⁰⁾ for these parameters have similar shape as can be seen in first three panels on the left of Fig. 4. Now, if we compare the graphs on all of the left panels of Figs. 2, 3, and 4, we see that although the (approximate) true covariance is known in all of these cases, there is a difference in the shapes of the probability distributions for ROI⁽⁰⁾. It can be clearly seen that in the experiments for the left panels of Fig. 2 the probability density function curves are approximately symmetric about their axis of symmetry, in which case the sample is exactly drawn from a Gaussian distribution. As a comparison, we see that for tests using small ensemble sizes, which is for the left panels of Figs. 3 and 4, the curves do not have symmetry at all. This is probably because the distributions of state variables in the Lorenz-96 system are not close to Gaussian. For $N = 21, T_1 = 50, T_2 = 10$ (red curve in Fig. 3, $N = 21$), the curve even has two peaks. In this case it would be hard for us to determine a good localization radius even if we know the true covariance. For $T_1 = 500, T_2 = 100$ (the black curves in Fig. 4), the behavior of ROI⁽⁰⁾ becomes close to that when $T_1 = 50, T_2 = 100$ for all ensemble sizes. This might be because after running the model for 100 time steps without data assimilation, the distribution of state variables become close to the ‘‘climatological distribution.’’ But as we compare the results in this case with the numerical results in section 4b, we see that ROI around 11 does not give optimal RMSE for the long time real implementation. However, the curves on the right panels of Figs. 3 and 4 are more symmetric. This might be because our probabilistic method still assumes Gaussian distribution and we do see in the cases $T_1 = 50$ or 500, $T_2 = 1$ or 10, ROI⁽¹⁾ would result in smaller RMSE according to numerical results in section 4b, hence, it is a better value for real-data implementation.

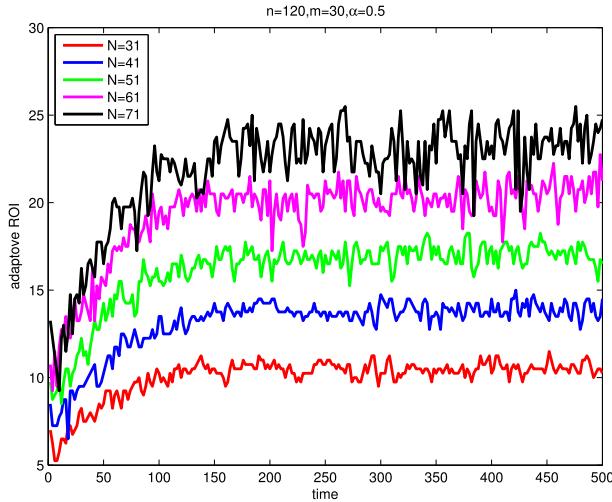


FIG. 5. Adaptive ROI from algorithm A1 as a function of time for different ensemble sizes. We can see that for larger ensemble size the resulting $\text{ROI}^{(1)}$ also gets larger. This experiment is done in Lorenz-96 system.

b. Experiments that use algorithm A1 to do adaptive localization in Lorenz-96 system

Now we implement algorithm A1 in the Lorenz-96 system. As mentioned before, the Lorenz-96 system configured here has $n = 120$ state variables and is integrated using the Runge–Kutta fourth-order scheme. The observations are uniformly distributed and lie on some of the grid points. The number of observations m and ensemble size N are different for each case. To handle filter divergence, we use a relaxation coefficient $\alpha = 0.5$ (Zhang et al. 2004). The inflation method is necessary here to prevent the model from blowing up. The choice of $\alpha = 0.5$ is not tuned. We also did the same experiment for $\alpha = 0.25$ and the results are similar. The choice of α can slightly influence the resulting $\text{ROI}^{(1)}$, but the change is ignorable at least for this experiment, so we only show the results for $\alpha = 0.5$. At every time step when observations are available, we first compute $\text{ROI}^{(1)}$ using algorithm A1, then we use $\text{ROI}^{(1)}$ to be the localization radius before assimilating the observations.

1) SENSITIVITY TO ENSEMBLE SIZE

In Fig. 5 we plot the curves of $\text{ROI}^{(1)}$ as a function of time for different ensemble sizes. In general we see that the resulting localization radius increases with the ensemble size, which is consistent with the results in section 2 where the true covariance is known.

2) COMPARISON OF RMSE FOR DIFFERENT OBSERVATIONAL DENSITY

In this subsection, we fix the ensemble size at $N = 61$ and plot the spatial RMSE of the ensemble mean as

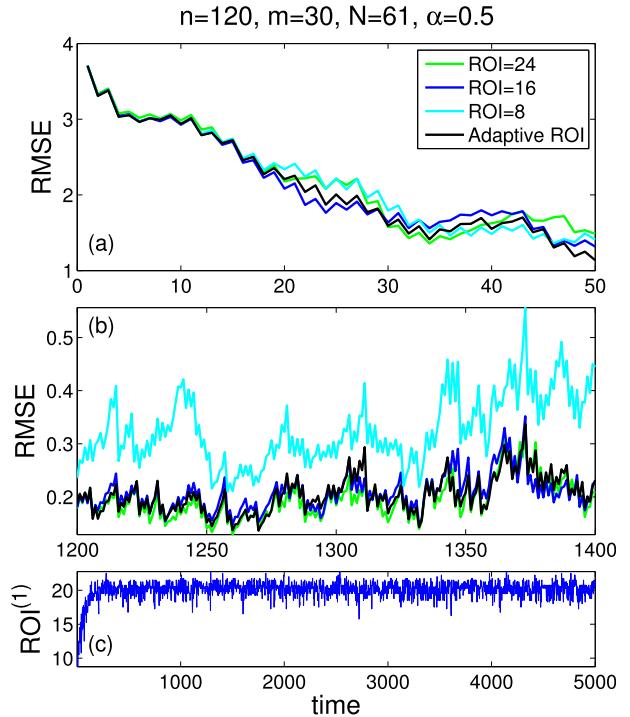


FIG. 6. Comparison of spatial RMSE of ensemble mean using adaptive ROI with that using fixed $\text{ROI} = 8, 16,$ and 24 in the Lorenz-96 system. We can see that the $\text{ROI}^{(1)}$ converges to around 20 after a few time steps. The RMSE for adaptive $\text{ROI}^{(1)}$ is comparable with that for optimal $\text{ROI} = 24$. The bottom plot shows how $\text{ROI}^{(1)}$ evolves with time.

a function of time for times steps between 1 and 50, 1000 and 1200, and 1200 and 1400. These values are plotted for $m = 30, 60,$ and $120,$ in Figs. 6, 7, and 8, respectively, along with $\text{ROI}^{(1)}$ as a function of time. While the larger ROI tends to result in smaller RMSEs, the solution can be unstable when the observations are sparse as indicated by the missing red curve in Fig. 6. The optimal choice of ROI must lower the RMSE, while maintaining a stable solution. In Table 6 we show the temporal mean of RMSE from time step 1000 to time step 5000 for different number of observations and fixed/adaptive ROI. When $m = 30,$ the temporal mean of RMSE for the green, blue, cyan, and black curve is 0.213, 0.2274, 0.4295, and 0.2296, respectively. We find that the $\text{ROI}^{(1)}$ value becomes stable after a few steps of data assimilation and lies around 20.2, and the RMSEs of the black, blue, and green curves are indistinguishable. In this case the adaptive method almost gives the best ROI. On the other hand, we found that the resulting ROI is smaller when the observation density increases. This is likely because assimilating denser observations may cause the ensemble correlation to be more narrowly supported, which then leads to a smaller localization radius by this algorithm. Our

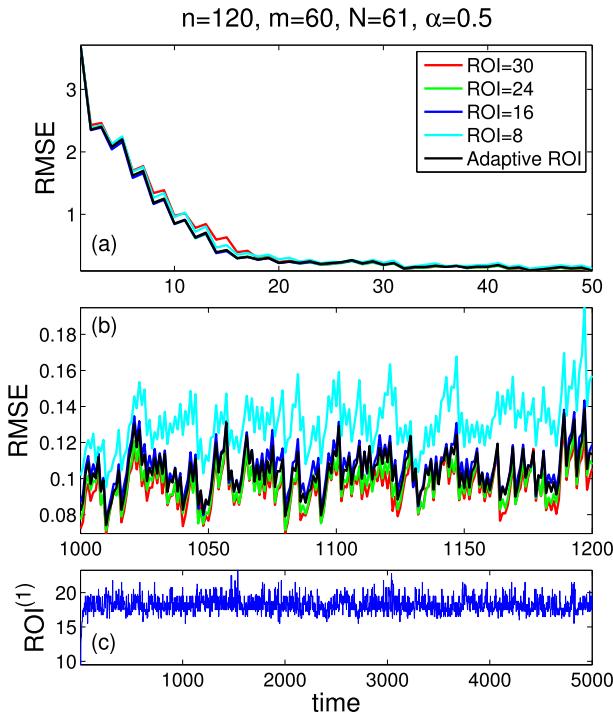


FIG. 7. Comparison of spatial RMSE of ensemble mean using adaptive ROI with that using fixed ROI = 8, 16, . . . , 30 in the Lorenz-96 system. We can see that the $ROI^{(1)}$ converges to around 18 after a few time steps. The RMSE for adaptive $ROI^{(1)}$ is slightly larger than that for optimal ROI = 30. The bottom plot shows how $ROI^{(1)}$ evolves with time.

speculation is based on Zhang et al. (2006, 729–730) and Daley and Menard (1993, see their Fig. 2), which is often called the “whitening” of the analysis error spectrum in that larger-scale error will be more effectively reduced first with more observations assimilated. In the meantime it is known that with denser observations larger localization radius can be applied to improve the mean update while also maintaining the stability of the solution. When $m = 60$, the temporal mean of the red, green, blue, cyan, and black curves is 0.1038, 0.1061, 0.1138, 0.1381, and 0.1112. The temporal mean of $ROI^{(1)}$ in this case is 18.2209. When $m = 120$, the temporal mean of the red, green, blue, cyan, and black curves is 0.0652, 0.0689, 0.0718, 0.0854, and 0.0713 while the temporal mean of $ROI^{(1)}$ is 16.5928. In Fig. 8, we see that the blue and black curves are nearly identical,

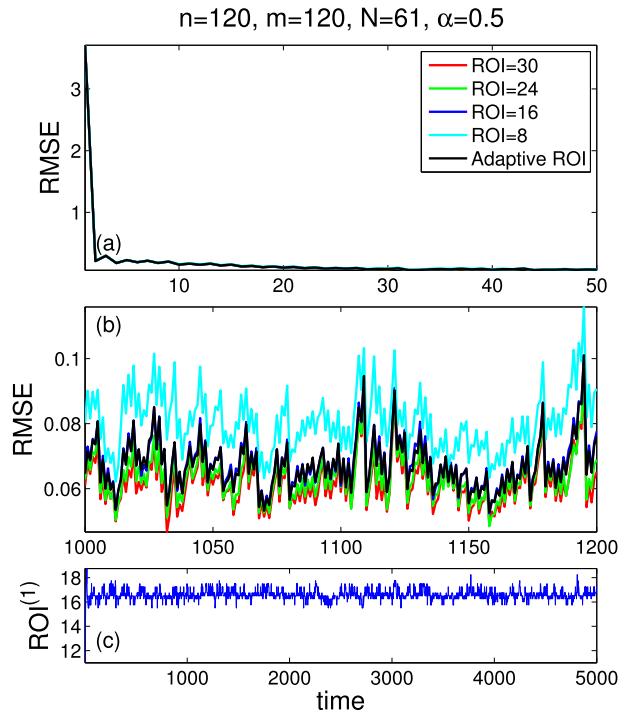


FIG. 8. Comparison of spatial RMSE of ensemble mean using adaptive ROI with that using fixed ROI = 8, 16, . . . , 30 in the Lorenz-96 system. We can see that the $ROI^{(1)}$ converges to around 16 after a few time steps. The RMSE for adaptive $ROI^{(1)}$ is slightly larger than that for optimal ROI = 30. The bottom plot shows how $ROI^{(1)}$ evolves with time.

because the temporal mean $ROI^{(1)}$ is approximately 16. And from the graph we can easily see that the red curve is consistently better than the black curve implying that there is still room to reduce the RMSE by enlarging the ROI. This suggests that more work needs to be done to take observation density into account so the localization radius determined by this algorithm can be more efficient. On the other hand, comparing the graphs in the top panels we see that with more observations the RMSE reduces more quickly and it is hard to distinguish the curves for the first 50 time steps in Figs. 7 and 8.

3) CURVE OF F VALUE

The value of the cost function F for each observation in the $m = 30$ cases is plotted in the top of Fig. 9 as a function of ROI for time steps 2, 120, and 1020. The

TABLE 6. The temporal RMSE of the ensemble mean from time steps 1000 to 5000 in the Lorenz-96 system for different observation densities and fixed/adaptive ROI. NaN refers to missing data caused by frequent model breakdown.

	Red (ROI = 30)	Green (ROI = 24)	Blue (ROI = 16)	Cyan (ROI = 8)	Black (adaptive ROI)
$m = 30$	NaN	0.213	0.2274	0.4295	0.2296
$m = 60$	0.1038	0.1061	0.1138	0.1381	0.1112
$m = 120$	0.0652	0.0689	0.0718	0.0854	0.0713

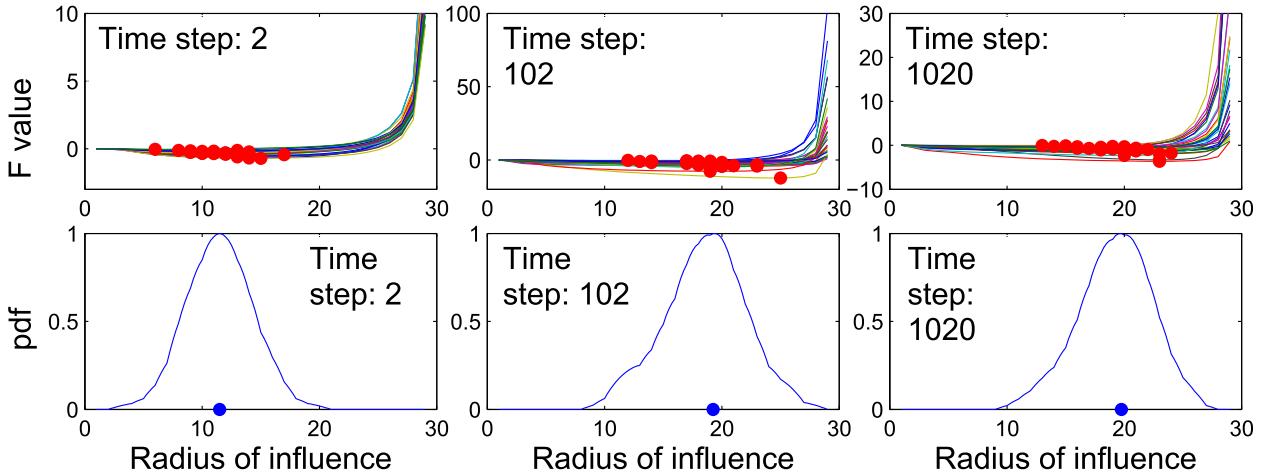


FIG. 9. (top) F -value curves [defined in Eq. (11)] for different observations and different time steps, and (bottom) the estimated (rescaled) probability density of ROI. This experiment is done in the Lorenz-96 system. Since we have $m = 30$ observations, we have 30 curves in each panel at the top. The red dots are where the minimum of F value are taken for each observation. The blue dots are the maximum likelihood estimate of $ROI^{(1)}$ at each time step. Hence, the blue dots are the output $ROI^{(1)}$ of algorithm A1 at each time step.

ensemble size $N = 61$. The red dots indicate where F is a minimum for each observation. In the bottom panels, the curves are the (rescaled) estimated probability density of all possible ROIs. The blue dot shows where the maximum likelihood is taken, hence it is also the final $ROI^{(1)}$ we use for localization at each time step. The curves in the top panels are almost convex, making finding the minimum value (red dots) of F for each observation not a trivial problem. This is also the consequence of adding the penalty term to the cost function. The bottom panels show that the optimal $ROI^{(1)}$ for each observation is distributed in a way that the probability density function of $ROI^{(1)}$ only has one peak (i.e., the maximum likelihood estimate is clearly well defined). If the density function has two peaks, then it is not easy to determine a unique localization radius. In that case, one can try dividing observations into groups so for each group of observations the probability density function of $ROI^{(1)}$ is better behaved. This aspect of the algorithm needs further exploration in the future.

5. Discussion and summary

In this article we first presented a cost function approach to analyze the sampling error issue in the case where the true covariance is known. Then we generalized this method to the case where the true covariance is not known. We presented a probabilistic approach to determine the localization radius adaptively when serial ensemble square root filters are used to assimilate uncorrelated observations. The advantage of this method is that it uses the information from the ensemble members

and observations only within a single assimilation window. Further the computational cost of this algorithm is small and its performance in the Lorenz-96 system is promising. We compared the results from this probabilistic method with that from the deterministic method in the case that the true covariance is known. The results show that the probabilistic method gives more useful radius of influence for long time implementation than the deterministic method. As a side result, we find that in order to determine a good localization radius, one needs to consider more than simply sampling error. In particular, the dynamical property of the model needs to be taken into account.

There are severe issues worth further discussion and investigation. First of all, the algorithm under discussion does not utilize any information regarding the observation density. As a consequence, the $ROI^{(1)}$ is not the optimal radius of influence when observations are dense. Second the algorithm in this article is based on serial ensemble square root Kalman filters and has the requirement that the observational operator H must be the restriction operator and that observations must be uncorrelated, which must be eliminated in the future work. Third the output of $ROI^{(0)}$ in the experiments in Lorenz-96 system, where the true covariance is approximated by 6000 ensemble members, is clearly not optimal according to the curves in Fig. 8, suggesting again that in order to get a good localization radius other system dynamics besides sampling error need to be considered. On the other hand the choice of a noninformative prior and the use of an empirical penalty term are not necessarily to be optimal as presented in this article. Deriving a useful mathematical formula for other choices of prior is a challenging problem.

There are several ways to make modifications in the algorithm in order to make use of more statistical tools. For example, we can generalize the concept of maximum likelihood as the “peaks of the density function.” More precisely speaking, if the probability density function of ROI⁽¹⁾ has more than one peak, it may be more wise to use the different peak values as the ROI for different observations rather than using only the peak value at the maximum likelihood estimate. One can also try incorporating this scheme with other methods. For example, one can consider taking observational impact into account to get a weighted maximum likelihood estimate. Future research is also warranted to extend the current one-dimensional study to 2D or 3D as well as for covariance across different physical variable (e.g., between temperature and winds).

Acknowledgments. We thank Zhibiao Zhao for his valuable suggestion on the choice of Jeffreys prior; Ningtao Wang and Cheng You for their valuable suggestion about Wishart distribution; Jeffrey Anderson for his encouraging us to publish this result; John Harlim, Jonathan Poterjoy, and the anonymous reviewers for their patient and careful reading and generous suggestion on how to make the math more clear. This work was supported in part by Office of Naval Research Grant N000140910526 and the National Science Foundation Grant ATM-0840651.

APPENDIX

Mathematical Derivation

a. Notation

For convenience the notation in this appendix is a little different from those in the main context. For example, \mathbf{B}^s in the appendix is not sample covariance, but $(N - 1)\mathbf{B}_{Loc}^s$; \mathbf{B} is actually \mathbf{B}_{Loc} ; the true regression coefficient is not denoted by r_i^t , but by r_i ; and n is actually short for n_{Loc} .

b. Review of setup

We now consider how to determine the localization radius when assimilating a single observation. Since we only assimilate one observation each time, for the convenience of notation, we always assume the observation is on the first grid point. The observation operator \mathbf{H} is a row vector in this case and more specifically $\mathbf{H} = (1, 0, \dots, 0)$.

Let r_i^s be the regression coefficient for the i th variable when using the pure sample covariance matrix. Let r_i be the regression coefficient for the i th variable when using the true covariance matrix. Then the regression coefficient for the i th variable when using localized sample covariance matrix is $r_i^s \rho_i$ where $\rho_i = GC(ROI, d_i)$.

We do not know what the true covariance is hence we think of the true covariance as a random matrix. Let \mathbf{B}^s be $N - 1$ times the sample covariance matrix. Let $p(\mathbf{B} | \mathbf{B}^s)$ be the probability density of the true covariance matrix given \mathbf{B}^s . Then

$$p(\mathbf{B} | \mathbf{B}^s) = \frac{p(\mathbf{B}, \mathbf{B}^s)}{p(\mathbf{B}^s)} \propto p(\mathbf{B}, \mathbf{B}^s) = p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}).$$

Let $\mathbf{B} = (b_{ij})$, $\mathbf{B}^s = (b_{ij}^s)$, \mathbf{R} be the observational covariance, $\mathbf{H} = (1, 0, \dots, 0)$.

Then

$$r_i^s = \frac{b_{1,i}^s / (N - 1)}{\mathbf{R} + b_{1,1}^s / (N - 1)} = \frac{b_{1,i}^s}{(N - 1)\mathbf{R} + b_{1,1}^s}, \quad \text{and}$$

$$r_i = \frac{b_{1,i}}{\mathbf{R} + b_{1,1}}.$$

Let $n = n_{Loc}$ be the number of state variables within distance ROI to the observation and N be the ensemble size. We require $N > n_{Loc}$.

We define functions F_1 and F_2 :

$$F_1(ROI) = \mathbf{C}(S) \int \sum_{i=2}^n [(\rho_i r_i^s - r_i)^2 - r_i^2] p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B},$$

and

$$F_2(ROI) = \mathbf{C}(S) \int \sum_{i=2}^n \rho_i^2 (r_i^s - r_i)^2 p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B},$$

where

$$\mathbf{C}(S) = \frac{1}{p(\mathbf{B}^s)}$$

is a constant depending only on the sample and n such that

$$\mathbf{C}(S) \int p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} = 1. \quad (A1)$$

We want to minimize the following:

$$F(ROI) = F_1(ROI) + F_2(ROI). \quad (A2)$$

In this draft we use Jeffreys prior: $p(\mathbf{B}) = 1/[\det(\mathbf{B})^{(n+1)/2}]$.

The notations are

$$d\mathbf{B} = db_{1,1} db_{1,2} \dots db_{1,n} db_{2,2} \dots db_{2,n} \dots db_{n,n}$$

and $\text{etr}(\mathbf{A}) = e^{\text{tr}(\mathbf{A})}$

(Muirhead 1982).

c. *Mathematical derivation*

1) FIRST WE COMPUTE $\mathbf{C}(S)$:

$$\begin{aligned} 1 &= \mathbf{C}(S) \int p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}) d\mathbf{B} = \frac{\mathbf{C}(S)}{2^{\ln(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{\det(\mathbf{B}^s)^{[(N-n-2)/2]}}{\det(\mathbf{B})^{(N-1)/2}} \text{etr}\left(-\frac{1}{2}\mathbf{B}^{-1}\mathbf{B}^s\right) \frac{1}{\det(\mathbf{B})^{(N+1)/2}} d\mathbf{B} \\ &= \frac{\mathbf{C}(S)}{2^{\ln(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{\det(\mathbf{B}^s)^{[(N-n-2)/2]}}{\det(\mathbf{B})^{(N+n)/2}} \text{etr}\left(-\frac{1}{2}\mathbf{B}^{-1}\mathbf{B}^s\right) d\mathbf{B}. \end{aligned}$$

Let $\mathbf{C}^s = (\mathbf{B}^s)^{-1}$ and $\mathbf{C} = \mathbf{B}^{-1}$

$$\begin{aligned} &= \frac{\mathbf{C}(S)\det(\mathbf{B}^s)^{[(N-n-2)/2]}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \\ &\quad \times \int \det(\mathbf{C})^{(n+N)/2} \text{etr}\left[-\frac{1}{2}\mathbf{C}(\mathbf{C}^s)^{-1}\right] d\mathbf{C}^{-1}. \end{aligned}$$

[By [Muirhead \(1982\)](#) theorem 2.1.8, $d\mathbf{C}^{-1} = \det(\mathbf{C})^{-n-1}d\mathbf{C}$.]

$$\begin{aligned} &= \frac{\mathbf{C}(S)\det(\mathbf{B}^s)^{[(N-n-2)/2]}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \\ &\quad \times \int \det(\mathbf{C})^{[(N-n-2)/2]} \text{etr}\left[-\frac{1}{2}\mathbf{C}(\mathbf{C}^s)^{-1}\right] d\mathbf{C} \end{aligned}$$

[By [Muirhead \(1982\)](#) theorem 2.1.11.]

$$\begin{aligned} &= \frac{\mathbf{C}(S)\det(\mathbf{B}^s)^{[(N-n-2)/2]}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} 2^{n(N-1)/2} \Gamma_n \\ &\quad \times \left(\frac{N-1}{2}\right) \det(\mathbf{C}^s)^{(N-1)/2} = \frac{\mathbf{C}(S)}{\det(\mathbf{B}^s)^{(N+1)/2}}. \end{aligned}$$

Hence,

lemma 1: $\mathbf{C}(S) = \det(\mathbf{B}^s)^{(n+1)/2}$.

2) COMPUTE $\mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}) d\mathbf{B}$

$$\begin{aligned} \mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}) d\mathbf{B} &= -\frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \\ &\quad \times \int \frac{2t_i}{1 + \mathbf{R}t_1^2} t_1^{N-n-1} \text{etr}\left\{-\frac{1}{2}\mathbf{T}^\Gamma \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\mathbf{B}}_1^{-1} \end{pmatrix} \mathbf{T}\mathbf{B}^s\right\} \det(\tilde{\mathbf{B}}_1)^{-(N+n)/2} d\mathbf{T} d\tilde{\mathbf{B}}_1. \end{aligned}$$

It is known that for $n \times n$ SPD matrix \mathbf{A} , $d\mathbf{A}^{-1} = \det(\mathbf{A})^{-(n+1)}d\mathbf{A}$ [see [Muirhead \(1982\)](#), theorem 2.1.8].

Therefore $d\tilde{\mathbf{B}}_1 = \det(\tilde{\mathbf{B}}_1^{-1})^{-n}d\tilde{\mathbf{B}}_1^{-1}$ and

$$\mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}) d\mathbf{B} = -\frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{2t_i}{1 + \mathbf{R}t_1^2} \text{etr}\left\{-\frac{1}{2}\mathbf{T}^\Gamma \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\mathbf{B}}_1^{-1} \end{pmatrix} \mathbf{T}\mathbf{B}^s\right\} t_1^{N-n-1} \det(\tilde{\mathbf{B}}_1^{-1})^{(N-n)/2} d\tilde{\mathbf{B}}_1^{-1} d\mathbf{T}.$$

$$\begin{aligned} \mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B})p(\mathbf{B}) d\mathbf{B} &= \frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \\ &\quad \times \int \frac{b_{1i}}{\mathbf{R} + b_{11}} \det(\mathbf{B})^{-(N+n)/2} \text{etr}\left(-\frac{1}{2}\mathbf{B}^{-1}\mathbf{B}^s\right) d\mathbf{B}. \end{aligned}$$

Let

$$\mathbf{T} = \begin{pmatrix} t_1 & 0 \\ \mathbf{t} & \mathbf{I}_{n-1} \end{pmatrix},$$

where \mathbf{I}_{n-1} is the $(n-1) \times (n-1)$ identity matrix, 0 is the zero row vector, and $\mathbf{t} = (t_2, \dots, t_n)^\top$ is a column vector whose entries are defined by $t_i = -(b_{1i}/b_{11})$ for $2 \leq i \leq n$. We define $d\mathbf{T} = dt_1, \dots, dt_n$ and $d\mathbf{t} = dt_2, \dots, dt_n$. Let $\tilde{\mathbf{B}}_1$ be a $(n-1) \times (n-1)$ matrix whose entries are defined by $(\tilde{\mathbf{B}}_1)_{ij} = b_{i+1,j+1} - b_{1,i+1}b_{1,j+1}/b_{11}$. Then it can be straightforwardly verified that

$$\mathbf{T}\mathbf{B}\mathbf{T}^\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\mathbf{B}}_1 \end{pmatrix}$$

and we have

lemma 2: $d\mathbf{B} = 2b_{11}^{n+(1/2)}d\mathbf{T}d\tilde{\mathbf{B}}_1 = 2t_1^{-1-2n}d\mathbf{T}d\tilde{\mathbf{B}}_1$.

Because of space limitations we merely remark that this can be proven by careful but straightforward computation, leaving to the interested reader the task of filling in the details.

It is easy to see that $\det(\mathbf{B}) = (1/t_1^2)\det(\tilde{\mathbf{B}}_1)$. So

Now define $\mathbf{B}_1 = \tilde{\mathbf{B}}_1^{-1}$. So

$$= -\frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2}\Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{2t_i}{1 + \mathbf{R}t_1^2} \text{etr}\left\{-\frac{1}{2}\mathbf{T}^\top \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B}_1 \end{pmatrix} \mathbf{T}\mathbf{B}^s\right\} t_1^{N-n-1} \det(\mathbf{B}_1)^{(N-n)/2} d\mathbf{B}_1 d\mathbf{T}.$$

Let the ensemble perturbation be $\mathbf{E} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \mathbf{E}_1 \end{pmatrix}$ and define $\boldsymbol{\delta}_i = \boldsymbol{\epsilon}_i + t_i\boldsymbol{\epsilon}_1$. Then by direct computation one can get

$$\begin{aligned} \text{tr}\left[\mathbf{T}^\top \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B}_1 \end{pmatrix} \mathbf{T}\mathbf{B}^s\right] &= \text{tr}\left[\mathbf{T}^\top \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B}_1 \end{pmatrix} \mathbf{T}\mathbf{E}\mathbf{E}^\top\right] \\ &= \text{tr}\left[\mathbf{E}^\top \mathbf{T}^\top \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B}_1 \end{pmatrix} \mathbf{T}\mathbf{E}\right] = (\boldsymbol{\epsilon}_1 t_1)^2 + \text{tr}\left[\begin{pmatrix} \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_n \end{pmatrix} \mathbf{B}_1 \begin{pmatrix} \boldsymbol{\delta}_2 & \dots & \boldsymbol{\delta}_n \end{pmatrix}\right]. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} &= \\ -\frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2}\Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{2t_i}{1 + \mathbf{R}t_1^2} t_1^{N-n-1} e^{-(1/2)t_1^2|\boldsymbol{\epsilon}_1|^2} \text{etr}\left[-\frac{1}{2}(\boldsymbol{\delta}_2^\top, \dots, \boldsymbol{\delta}_n^\top) \mathbf{B}_1 \begin{pmatrix} \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_n \end{pmatrix}\right] \det(\mathbf{B}_1)^{(N-n)/2} d\mathbf{B}_1 d\mathbf{T}. \end{aligned}$$

Let $\boldsymbol{\Delta}_t = \begin{pmatrix} \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_n \end{pmatrix} (\boldsymbol{\delta}_2^\top, \dots, \boldsymbol{\delta}_n^\top)$. Then

$$\begin{aligned} \mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} &= \\ -\frac{\det(\mathbf{B}^s)^{(N-1)/2}}{2^{n(N-1)/2}\Gamma_n\left(\frac{N-1}{2}\right)} \int \frac{2t_i}{1 + \mathbf{R}t_1^2} t_1^{N-n-1} e^{-(1/2)\boldsymbol{\epsilon}_1^\top t_1^2 \boldsymbol{\epsilon}_1} \text{etr}\left(-\frac{1}{2}\mathbf{B}_1 \boldsymbol{\Delta}_t\right) \det(\mathbf{B}_1)^{(N-n)/2} d\mathbf{B}_1 d\mathbf{T}. \end{aligned}$$

It follows from [Muirhead \(1982\)](#) theorem 2.1.11 that

$$\int \text{etr}\left(-\frac{1}{2}\mathbf{B}_1 \boldsymbol{\Delta}_t\right) \det(\mathbf{B}_1)^{(N-n)/2} d\mathbf{B}_1 = 2^{N(n-1)/2} \Gamma_{n-1}\left(\frac{N}{2}\right) \det(\boldsymbol{\Delta}_t)^{-N/2}.$$

Hence,

$$\mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} = -\frac{\det(\mathbf{B}^s)^{N-1/2} \Gamma_{n-1}\left(\frac{N}{2}\right)}{|\boldsymbol{\epsilon}_1|^{N-n-2} 2^{(N-n)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \left\{ \int_0^\infty \frac{2t_1^{N-n-1}}{|\boldsymbol{\epsilon}_1|^2 + \mathbf{R}t_1^2} e^{-(1/2)t_1^2} dt_1 \right\} \left\{ \int \frac{t_i}{\det(\boldsymbol{\Delta}_t)^{N/2}} dt \right\}.$$

Let

$$\mathbf{e} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_1^\top \\ \vdots \\ \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_n^\top \end{pmatrix}$$

and $\mathbf{s} = |\boldsymbol{\epsilon}_1| \mathbf{t} + \mathbf{e}/|\boldsymbol{\epsilon}_1|$, $(\mathbf{E}_0)_{ij} = \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^\top$, $e_i = \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_i^\top$, for $2 \leq i$ and $j \leq n$.

Then

$$\boldsymbol{\Delta}_t = \mathbf{s}\mathbf{s}^\top + \mathbf{E}_0 - \frac{\mathbf{e}\mathbf{e}^\top}{|\boldsymbol{\epsilon}_1|^2},$$

$$ds = \left| \det \frac{\partial(\mathbf{s})}{\partial(\mathbf{t})} \right| dt = \left| \det \frac{\partial(s_2, \dots, s_n)}{\partial(t_2, \dots, t_n)} \right| dt = |\boldsymbol{\epsilon}_1|^{n-1} dt,$$

and $d\mathbf{T} = dt_1 dt$.

Hence, it suffices to compute the following:

$$a_1 = \int_0^\infty e^{-(1/2)\boldsymbol{\epsilon}_1^\top t_1^2 \boldsymbol{\epsilon}_1} \frac{t_1^{N-n-1}}{1 + \mathbf{R}t_1^2} dt_1,$$

$$a_2 = \int \frac{t_i}{\det(\Delta_t)^{N/2}} dt,$$

$$a_3 = \int_0^\infty e^{-(1/2)|\epsilon_1^T t_1|^2} t_1^{N-n-1} dt_1.$$

Lemma 3: For $2l > n + 2$ and even integer n ,

$$\int_{-\infty}^\infty \dots \int_{-\infty}^\infty \frac{1}{(a^2 + u_1^2 + \dots + u_n^2)^l} du_1 \dots du_n = \frac{(2l - n - 2)!!}{(2l - 2)!!} (2\pi)^{n/2}$$

Lemma 4: For $2l > n + 2$, $\int_{-\infty}^\infty \dots \int_{-\infty}^\infty [u_i/(a^2 + u_1^2 + \dots + u_n^2)^l] du_1 \dots du_n = 0$.

Lemma 5: Let $\mathbf{I} \in \mathbb{R}^{n \times n}$ denote the identity matrix and $\mathbf{s} \in \mathbb{R}^n$ be a column vector. Then $\det(\mathbf{I} + \mathbf{ss}^T) = 1 + |\mathbf{s}|^2$.

Lemma 6:

$$a_3 = \Gamma\left(\frac{N-n}{2}\right) \frac{2^{(N-n-2)/2}}{|\epsilon_1|^{N-n}},$$

$$a_2 = -\frac{1}{|\epsilon_1|^{n-1}} \frac{1}{\det\left(\mathbf{E}_0 - \frac{\mathbf{ee}^T}{|\epsilon_1|^2}\right)^{N-1/2}} \frac{e_i}{|\epsilon_1|^2} \frac{(N-n-1)!!}{(N-2)!!} (2\pi)^{n-1/2},$$

where $\mathbf{E}_0 - (\mathbf{ee}^T/|\epsilon_1|^2) = \mathbf{V}^T \mathbf{D} \mathbf{V}$ is the singular value decomposition, and $(\mathbf{V}^T)_i$ is the i th row of the matrix \mathbf{V}^T .

Proof: Recall that $\Delta_t = \mathbf{ss}^T + \mathbf{E}_0 - (\mathbf{ee}^T/|\epsilon_1|^2)$ and $\mathbf{s} = |\epsilon_1| \mathbf{t} + (\mathbf{e}/|\epsilon_1|)$:

$$\begin{aligned} a_2 &= \int \frac{t_i}{\det(\Delta_t)^{N/2}} dt \\ &= \int \frac{s_i}{|\epsilon_1| \det(\mathbf{ss}^T + \mathbf{V}^T \mathbf{D} \mathbf{V})^{N/2}} - \frac{e_i}{|\epsilon_1|^2 \det(\mathbf{ss}^T + \mathbf{V}^T \mathbf{D} \mathbf{V})^{N/2}} dt \\ &= \frac{1}{|\epsilon_1|^n} \int \frac{s_i}{\det(\mathbf{ss}^T + \mathbf{V}^T \mathbf{D} \mathbf{V})^{N/2}} ds - \frac{1}{|\epsilon_1|^{n+1}} \int \frac{e_i}{\det(\mathbf{ss}^T + \mathbf{V}^T \mathbf{D} \mathbf{V})^{N/2}} ds. \end{aligned}$$

Let $\beta = \sqrt{\mathbf{D}^{-1}} \mathbf{Vs}$ as before, then $\mathbf{s} = \mathbf{V}^T \sqrt{\mathbf{D}} \beta$ and $ds = \det(\mathbf{D})^{1/2} d\beta$:

$$\begin{aligned} &= \frac{1}{|\epsilon_1|^n \det(\mathbf{D})^{(N-1)/2}} \int \frac{(\mathbf{V}^T)_i \sqrt{\mathbf{D}} \beta}{\det(\beta \beta^T + \mathbf{I})^{N/2}} d\beta - \frac{e_i}{|\epsilon_1|^{n+1} \det(\mathbf{D})^{(N-1)/2}} \int \frac{1}{\det(\mathbf{I} + \beta \beta^T)^{N/2}} d\beta \\ &= -\frac{1}{|\epsilon_1|^{n-1} \det(\mathbf{D})^{(N-1)/2}} \frac{e_i}{|\epsilon_1|^2} \frac{(N-n-1)!!}{(N-2)!!} (2\pi)^{(n-1)/2}, \end{aligned}$$

where a_3 is essentially a Gamma function. One can easily get the desired expression of a_3 by substitution.

$$\Delta = \mathbf{E}_0 - \frac{\mathbf{ee}^T}{|\epsilon_1|^2}$$

Let

then

$$\begin{aligned} \mathbf{C}(S) &\int r_i p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} \\ &= \frac{\det(\mathbf{B}^s)^{(N-1)/2} \Gamma_{n-1}\left(\frac{N}{2}\right)}{|\epsilon_1|^{N-n-2} 2^{(N-n)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \left\{ \int_0^\infty \frac{2t_1^{N-n-1}}{|\epsilon_1|^2 + \mathbf{R}t_1^2} e^{-(1/2)t_1^2} dt_1 \right\} \frac{1}{|\epsilon_1|^{n-1}} \frac{1}{\det(\Delta)^{(N-1)/2}} \frac{e_i}{|\epsilon_1|^2} \frac{(N-n-1)!!}{(N-2)!!} (2\pi)^{(n-1)/2}. \end{aligned}$$

Recall the following (well known) identities about multivariate gamma functions (assuming n, N are both odd numbers):

$$\Gamma_n\left(\frac{N-1}{2}\right) = \pi^{n(n-1)/4} \Gamma\left(\frac{N-1}{2}\right) \Gamma\left(\frac{N-2}{2}\right) \dots \Gamma\left(\frac{N-n}{2}\right),$$

$$\Gamma_{n-1}\left(\frac{N}{2}\right) = \pi^{[(n-2)(n-1)]/4} \Gamma\left(\frac{N}{2}\right) \Gamma\left(\frac{N-1}{2}\right) \dots \Gamma\left(\frac{N-n+2}{2}\right),$$

$$\Gamma\left(\frac{N-n+1}{2}\right) \Gamma\left(\frac{N-n}{2}\right) = 2^{1+n-N} \sqrt{\pi} \Gamma(N-n),$$

$$\Gamma\left(\frac{N}{2}\right) = 2^{(1-N)/2} \sqrt{\pi} \Gamma(N-2)!!$$

It is not hard to check that

$$\det(\mathbf{B}^s) = \det(\mathbf{\Delta}) |\boldsymbol{\epsilon}_1|^2.$$

Hence (combining all the above identities), we have lemma 7:

$$\begin{aligned} \mathbf{C}(S) \int r_i p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} \\ = r_i^s \frac{\int_0^\infty \frac{t_1^{N-n-1} e^{-(1/2)t_1^2} dt_1}{|\boldsymbol{\epsilon}_1|^2 + \mathbf{R}t_1^2}}{\int_0^\infty \frac{t_1^{N-n-1} e^{-(1/2)t_1^2} dt_1}{|\boldsymbol{\epsilon}_1|^2 + (N-1)\mathbf{R}}} \end{aligned} \quad (\text{A3})$$

3) COMPUTE $\mathbf{C}(S) \int r_i^2 p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B}$:

By similar derivation we can get

$$\begin{aligned} \mathbf{C}(S) \int r_i^2 p(\mathbf{B}^s | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} \\ = \frac{\det(\mathbf{B}^s)^{(N-1)/2} \Gamma_{n-1}\left(\frac{N}{2}\right)}{|\boldsymbol{\epsilon}_1|^{N-n-4} 2^{(N-n)/2} \Gamma_n\left(\frac{N-1}{2}\right)} \\ \times \left\{ \int_0^\infty \frac{2t_1^{N-n-1}}{(|\boldsymbol{\epsilon}_1|^2 + \mathbf{R}t_1^2)^2} e^{-(1/2)t_1^2} dt_1 \right\} \left\{ \int \frac{t_i^2}{\det(\mathbf{\Delta}_i)^{N/2}} dt \right\}. \end{aligned}$$

Hence, it suffices to compute

$$\int \frac{t_i^2}{\det(\mathbf{\Delta}_i)^{N/2}} dt.$$

Lemma 8:

$$\frac{dt}{\det(\mathbf{\Delta}_i)^{N/2}} = \frac{d\boldsymbol{\beta}}{\det(\mathbf{\Delta})^{(N-1)/2} (1 + |\boldsymbol{\beta}|^2)^{N/2} |\boldsymbol{\epsilon}_1|^{n-1}}.$$

It is easy to see that by symmetry when $i \neq j$

$$\int \frac{\beta_i \beta_j}{1 + |\boldsymbol{\beta}|^2} d\boldsymbol{\beta} = \int \frac{\beta_i}{1 + |\boldsymbol{\beta}|^2} d\boldsymbol{\beta} = 0,$$

$$\int \frac{\beta_i^2}{1 + |\boldsymbol{\beta}|^2} d\boldsymbol{\beta} = \int \frac{\beta_j^2}{1 + |\boldsymbol{\beta}|^2} d\boldsymbol{\beta}.$$

On the other hand,

$$\begin{aligned} t_i^2 = \frac{1}{|\boldsymbol{\epsilon}_1|^2} \left\{ \sum_{j=1}^{n-1} (\mathbf{V}^T)_{ij}^2 d_{jj} \beta_j^2 \right\} + \left\{ \sum_{i \neq j} (\dots) \beta_i \beta_j \right\} \\ + \left\{ \sum_{j=1}^{n-1} (\dots) \beta_j \right\} + \frac{e_i^2}{|\boldsymbol{\epsilon}_1|^4}. \end{aligned}$$

Hence, one can derive

$$\begin{aligned} \int \frac{t_i^2}{\det(\mathbf{\Delta}_i)^{N/2}} dt \\ = \frac{1}{\det(\mathbf{\Delta})^{(N-1)/2} |\boldsymbol{\epsilon}_1|^{n-1}} \left[\frac{\mathbf{\Delta}_{ii}}{|\boldsymbol{\epsilon}_1|^2} \int \frac{\beta_1^2}{(1 + |\boldsymbol{\beta}|^2)^{N/2}} d\boldsymbol{\beta} \right. \\ \left. + \frac{e_i^2}{|\boldsymbol{\epsilon}_1|^4} \int \frac{1}{(1 + |\boldsymbol{\beta}|^2)^{N/2}} d\boldsymbol{\beta} \right]. \end{aligned}$$

Lemma 9: For odd numbers $n, 2k$:

$$\begin{aligned} a_5 := \int \frac{x_1^2}{(a^2 + x_1^2 + x_2^2 + \dots + x_{n-1}^2)^k} dx_1 dx_2 \dots dx_{n-1} \\ = \frac{(2k - n - 3)!!}{(2k - 2)!!} \pi^{(n-1)/2} 2^{(n-1)/2} \frac{1}{a^{2k-n-1}}. \end{aligned}$$

And similarly, we can derive

Lemma 10:

$$\begin{aligned} \mathbf{C}(S) \int r_i^2 p(\mathbf{S} | \mathbf{B}) p(\mathbf{B}) d\mathbf{B} \\ = \left(\frac{\mathbf{\Delta}_{ii} |\boldsymbol{\epsilon}_1|^2}{N - n - 1} + e_i^2 \right) \frac{\int_0^\infty \frac{t_1^{N-n-1}}{(|\boldsymbol{\epsilon}_1|^2 + \mathbf{R}t_1^2)^2} e^{-(1/2)t_1^2} dt_1}{\int_0^\infty t_1^{N-n-1} e^{-(1/2)t_1^2} dt_1}. \end{aligned} \quad (\text{A4})$$

Combining all the results above, we have the following theorem:

Theorem 5.1:

$$F_1(\text{ROI}) = \sum_{i=2}^n \left\{ \rho_i (r_i^s)^2 - 2\rho_i r_i^s e_i \frac{\int_0^\infty \frac{e^{-(1/2)t_1^2} t_1^{N-n-1}}{\mathbf{R}t_1^2 + |\boldsymbol{\epsilon}_1|^2} dt_1}{\int_0^\infty e^{-(1/2)t_1^2} t_1^{N-n-1} dt_1} \right\}, \quad (\text{A5})$$

$$F_2(\text{ROI}) = \sum_{i=2}^n \rho_i^2 \left\{ (r_i^s)^2 - 2r_i^s e_i \frac{\int_0^\infty e^{-(1/2)t_1^2} \frac{t_1^{N-n-1}}{\mathbf{R}t_1^2 + |\boldsymbol{\epsilon}_1|^2} dt_1}{\int_0^\infty e^{-(1/2)t_1^2} t_1^{N-n-1} dt_1} \right. \\ \left. + \left(\frac{\Delta_{ii} |\boldsymbol{\epsilon}_1|^2}{N-n-1} + e_i^2 \right) \frac{\int_0^\infty e^{-(1/2)t_1^2} \frac{t_1^{N-n-1}}{(\mathbf{R}t_1^2 + |\boldsymbol{\epsilon}_1|^2)^2} dt_1}{\int_0^\infty e^{-(1/2)t_1^2} t_1^{N-n-1} dt_1} \right\}, \quad (\text{A6})$$

where Δ_{ii} is the i th diagonal element of $\boldsymbol{\Delta}$.

It is not hard to see that

$$r_i^s = \frac{e_i}{(N-1)\mathbf{R} + |\boldsymbol{\epsilon}_1|^2}.$$

And by doing the same substitution as in the computation of a_3 in the proof of lemma 7, it is not hard to see the equivalence between Eqs. (A5), (A6) and Eqs. (4), (5).

REFERENCES

- Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi:10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.
- , 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, doi:10.1175/MWR-D-11-00013.1.
- , and L. Lei, 2013: Empirical localization of observation impact in ensemble Kalman filters. *Mon. Wea. Rev.*, **141**, 4140–4153, doi:10.1175/MWR-D-12-00330.1.
- Bernardo, J. M., and A. F. Smith, 1994: *Bayesian Theory*. John Wiley and Sons, 608 pp.
- Bishop, C. H., and D. Hodyss, 2007: Flow adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Quart. J. Roy. Meteor. Soc.*, **133**, 2029–2044, doi:10.1002/qj.169.
- , 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, doi:10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.
- , D. Hodyss, P. Steinle, H. Sims, A. M. Clayton, A. C. Lorenc, D. M. Barker, and M. Buehner, 2011: Efficient ensemble covariance localization in variational data assimilation. *Mon. Wea. Rev.*, **139**, 573–580, doi:10.1175/2010MWR3405.1.
- Daley, R., and R. Menard, 1993: Spectral characteristics of Kalman filter systems for atmospheric data assimilation. *Mon. Wea. Rev.*, **121**, 1554–1565, doi:10.1175/1520-0493(1993)121<1554:SCOKFS>2.0.CO;2.
- Evensen, G., 1994: Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, doi:10.1029/94JC00572.
- Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, doi:10.1002/qj.49712555417.
- Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter–3D variational analysis scheme. *Mon. Wea. Rev.*, **128**, 2905–2919, doi:10.1175/1520-0493(2000)128<2905: AHEKFF>2.0.CO;2.
- , J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, doi:10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.
- Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137, doi:10.1175/1520-0493(2001)129<0123: ASEKFF>2.0.CO;2.
- Jun, M., I. Szunyogh, M. G. Genton, F. Zhang, and C. H. Bishop, 2011: A statistical investigation of sensitivity of ensemble-based Kalman filters to covariance filtering. *Mon. Wea. Rev.*, **139**, 3036–3051, doi:10.1175/2011MWR3577.1.
- Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME J. Fluids Eng.*, **82D**, 35–45, doi:10.1115/1.3662552.
- Lei, L., and J. Anderson, 2014: Comparisons of empirical localization techniques for serial ensemble Kalman filters in a simple atmospheric general circulation model. *Mon. Wea. Rev.*, **142**, 739–754, doi:10.1175/MWR-D-13-00152.1.
- Lorenc, A., 2003: The potential of the ensemble Kalman filter for NWP—A comparison with 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **129**, 3183–3203, doi:10.1256/qj.02.132.
- Lorenz, E. N., 2006: Predictability—A problem partly solved. *Predictability of Weather and Climate*, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 40–58.
- Muirhead, R. J., 1982: *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, 673 pp.
- Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428, doi:10.1111/j.1600-0870.2004.00076.x.
- Snyder, C., and F. Zhang, 2003: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **131**, 1663–1677, doi:10.1175/2555.1.
- Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485–1490, doi:10.1175/1520-0493(2003)131<1485: ESRF>2.0.CO;2.
- Wang, X., D. M. Barker, C. Snyder, and T. M. Hamill, 2008a: A hybrid ETKF–3DVar data assimilation scheme for the WRF model. Part I: Observing system simulation experiment. *Mon. Wea. Rev.*, **136**, 5116–5131, doi:10.1175/2008MWR2444.1.
- , —, —, and —, 2008b: A hybrid ETKF–3DVar data assimilation scheme for the WRF model. Part II: Real observation experiments. *Mon. Wea. Rev.*, **136**, 5132–5147, doi:10.1175/2008MWR2445.1.
- Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, doi:10.1175/1520-0493(2002)130<1913: EDAWPO>2.0.CO;2.

- Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132**, 1238–1253, doi:[10.1175/1520-0493\(2004\)132<1238:IOIEAO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2).
- , Z. Meng, and A. Aksoy, 2006: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part I: Perfect model experiments. *Mon. Wea. Rev.*, **134**, 722–736, doi:[10.1175/MWR3101.1](https://doi.org/10.1175/MWR3101.1).
- , M. Zhang, and J. A. Hansen, 2009: Coupling ensemble Kalman filter with four-dimensional variational data assimilation. *Adv. Atmos. Sci.*, **26**, 1–8, doi:[10.1007/s00376-009-0001-8](https://doi.org/10.1007/s00376-009-0001-8).
- , —, and J. Poterjoy, 2013: E3DVar: Coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Mon. Wea. Rev.*, **141**, 900–917, doi:[10.1175/MWR-D-12-00075.1](https://doi.org/10.1175/MWR-D-12-00075.1).