PLOS ONE

RESEARCH ARTICLE

An approach to localization for ensemble-based data assimilation

Bin Wang[1,2,3], Juanjuan Liu[1,3]*, Li Liu[2], Shiming Xu[2], Wenyu Huang[2]

1 LASG, Institute of Atmospheric Physics, Beijing, China, **2** Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, China, **3** University of Chinese Academy of Sciences, Beijing, China

* ljjxgg@mail.iap.ac.cn

Citation: Wang B, Liu J, Liu L, Xu S, Huang W (2018) An approach to localization for ensemble-based data assimilation. PLoS ONE 13(1): e0191088. https://doi.org/10.1371/journal.pone.0191088

Editor: João Miguel Dias, Universidade de Aveiro, PORTUGAL

Received: October 14, 2016

Accepted: December 10, 2017

Published: January 19, 2018

Data Availability Statement: All relevant data are within the paper.

Funding: BW acknowledges the National Natural Science Foundation of China and the National Basic Research Program of China (973 Program) under Grant No. 91530204 and Grant No. 2014CB441302, respectively. JL is grateful to the National Natural Science Foundation of China (No. 91737307), the China Meteorological Administration for the R&D Special Fund for Public Welfare Industry (Meteorology) under Grant No. GYHY(QX)201406015. This work was also

Abstract

Localization techniques are commonly used in ensemble-based data assimilation (e.g., the Ensemble Kalman Filter (EnKF) method) because of insufficient ensemble samples. They can effectively ameliorate the spurious long-range correlations between the background and observations. However, localization is very expensive when the problem to be solved is of high dimension (say 10^6 or higher) for assimilating observations simultaneously. To reduce the cost of localization for high-dimension problems, an approach is proposed in this paper, which approximately expands the correlation function of the localization matrix using a limited number of principal eigenvectors so that the Schür product between the localization matrix and a high-dimension covariance matrix is reduced to the sum of a series of Schür products between two simple vectors. These eigenvectors are actually the sine functions with different periods and phases. Numerical experiments show that when the number of principal eigenvectors used reaches 20, the approximate expansion of the correlation function is very close to the exact one in the one-dimensional (1D) and two-dimensional (2D) cases. The new approach is then applied to localization in the EnKF method, and its performance is evaluated in assimilation-cycle experiments with the Lorenz-96 model and single assimilation experiments using a barotropic shallow water model. The results suggest that the approach is feasible in providing comparable assimilation analysis with far less cost.

Introduction

The statistical accuracy of background error is extremely important for any data assimilation scheme, and the background error covariance matrix (the **B** matrix, hereinafter) is often estimated from ensembles[1–4]. However, for practical applications in the ocean or atmosphere, current computational resources limit the ensemble size, which is often much smaller than both the dimension of the model and the number of observations. A small ensemble size is very likely to introduce sampling errors, leading to (a) underestimation of ensemble spread and (b) spurious correlations over long distances or between variables known to be uncorrelated [5]. Common approaches to reduce sampling errors in ensemble-based data assimilation (EDA) include covariance inflation[6–10] and localization[2,3,9,11–26].

In covariance inflation algorithms, the prior ensemble state covariance is increased by linearly inflating each scalar component of the state vector before assimilating observations

supported by the National Basic Research Program of China Grant No. 2015CB954102. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

[5,9,27]. By reducing the underestimation of the **B** matrix, covariance inflation plays an important role in preventing filter divergence of EDA. In addition, a proper localization of the estimated **B** matrix is needed to reduce spurious non-zero long-range correlations in the **B** matrix and to improve its rank deficiency, which allow ensemble-based assimilation schemes with an ensemble size fewer than 100 members to work properly with realistic atmosphere and ocean models [9,25].

Localization is usually implemented as a Schür product between the ensemble-based **B** matrix and a correlation matrix of which the elements are calculated according to a correlation function with respect to their coordinates. A commonly-used correlation function is the Quasi-Gaussian compactly supported form proposed by Gaspari and Cohn [28] (referred to simply as the GC localization or the correlation function, hereinafter). However, given that the optimal localization is likely to depend on the ensemble configuration (e.g., ensemble size, observation types), a comprehensive tuning of localization is needed in practice. In order to avoid the challenge of tuning the localization parameters [5, 25], adaptive localization functions have been proposed [9,17]. As noted in Fertig et al. [19], spatial localization is still difficult when assimilating satellite observations. To tackle this problem, Fertig et al. [19] updated the state at a given location through assimilating satellite observations that are strongly correlated to the model state there. In addition, Miyoshi and Sato [29] and Campbell et al.[30] explored localization functions for satellite radiances. In Miyoshi and Sato [29], the normalized sensitivity function of satellite sensors was used as the localization weights, whereas in Campbell et al.[30] forward operators performing weighted averages of a large number of state variables were applied. Zhu et al. [24] also proposed a localization scheme to use non-local observations; their basic idea is similar to that of Liu et al. [21].

Despite the differences among various localization approaches, the computational cost of localization algorithms is always an important issue as more and more observations are used. In the following, we provide a simple comparison between the computational costs with and without localization, based on the ensemble Kalman Filter (EnKF) with simultaneous treatment for assimilating observations.

Let m_x, m_y and n be the number of control variables, the number of observations and the ensemble size, respectively. In the EnKF approach, the forecast error covariance matrix $\mathbf{P}^f$ is calculated as follows:

$$\begin{cases} \mathbf{P}^f \approx \mathbf{b}\mathbf{b}^T \\ \mathbf{b} = \mathbf{b}_{m_x \times n} = \dfrac{1}{\sqrt{n-1}}\left(\mathbf{x}_1 - \overline{\mathbf{x}}, \mathbf{x}_2 - \overline{\mathbf{x}}, \ldots, \mathbf{x}_n - \overline{\mathbf{x}}\right). \\ \overline{\mathbf{x}} = \dfrac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n) \end{cases} \tag{1}$$

The gain matrix is then:

$$\mathbf{K}_{\{full\}} = \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}, \tag{2}$$

where **H** is the observation operator ($m_y \times m_x$ matrix) and **R** is the observational error covariance ($m_y \times m_y$ matrix). Since n is usually smaller than m_x, the sample covariance matrix is rank deficient. A Schür product is then implemented between the covariance matrix and a correlation matrix to increase the rank through removing spurious long-range correlations. Then, the Kalman gain is written as

$$\mathbf{K}_{\{loc\}} = (\boldsymbol{\rho}_{m_x \times m_x} \circ \mathbf{P}^f)\mathbf{H}^T(\mathbf{H}(\boldsymbol{\rho}_{m_x \times m_x} \circ \mathbf{P}^f)\mathbf{H}^T + \mathbf{R})^{-1}, \tag{3}$$

where $\boldsymbol{\rho}_{m_x \times m_x}$ is a compactly supported correlation matrix in which each column represents

spatial correlations at a given model gridpoint. This method is referred to as model spatial localization. Following Houtekamer and Mitchell [13], localization has also been applied in observation space, leading to the following Kalman gain:

$$\mathbf{K}_{\{loc\}} = \boldsymbol{\rho}_{m_x \times m_y} \circ (\mathbf{P}^f \mathbf{H}^T)(\boldsymbol{\rho}_{m_y \times m_y} \circ (\mathbf{H}\mathbf{P}^f \mathbf{H}^T) + \mathbf{R})^{-1}. \quad (4)$$

Since

$$\mathbf{Hb} \approx \mathbf{P}^{\mathbf{y}}_{m_y \times n} = \frac{1}{\sqrt{n-1}}(\mathbf{y}_1 - \bar{\boldsymbol{y}}, \mathbf{y}_2 - \bar{\boldsymbol{y}}, \ldots, \mathbf{y}_n - \bar{\boldsymbol{y}}); \bar{\boldsymbol{y}} = \frac{1}{\boldsymbol{n}}\sum_{i=1}^{n} \mathbf{y}_i, \quad (5)$$

where $\mathbf{P}^{\mathbf{y}}$ is a $m_y \times n$ matrix, Eqs (2) and (4) can be respectively expressed as

$$\mathbf{K}_{\{full\}} = \mathbf{K}_{m_x \times m_y} = \mathbf{b}_{m_x \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T (\mathbf{P}^{\boldsymbol{y}}_{m_y \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T + \mathbf{R}_{m_y \times m_y})^{-1}, \quad (6)$$

and

$$\mathbf{K}_{\{loc\}} = \mathbf{K}_{m_x \times m_y} = \boldsymbol{\rho}_{m_x \times m_y} \circ [\mathbf{b}_{m_x \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T](\boldsymbol{\rho}_{m_y \times m_y} \circ [\mathbf{P}^{\boldsymbol{y}}_{m_y \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T] + \mathbf{R}_{m_y \times m_y})^{-1}. \quad (7)$$

When the initial analysis increment $\boldsymbol{x}'_{m_x} = \mathbf{K}_{m_x \times m_y} \boldsymbol{y}'^{obs}_{m_y}$ is calculated using Eqs (6) and (7), respectively, their costs are quite different, where $\boldsymbol{y}'^{obs}_{m_y}$ is the m_y-dimension observation innovation vector. If we ignore the difference between the costs for calculating: $\boldsymbol{y}^r_{m_y} = (\mathbf{P}^{\boldsymbol{y}}_{m_y \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T + \mathbf{R}_{m_y \times m_y})^{-1} \boldsymbol{y}'^{obs}_{m_y}$ and $\boldsymbol{y}^{rl}_{m_y} = (\boldsymbol{\rho}_{m_y \times m_y} \circ [\mathbf{P}^{\boldsymbol{y}}_{m_y \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T] + \mathbf{R}_{m_y \times m_y})^{-1} \boldsymbol{y}'^{obs}_{m_y}$, which are relatively small, the calculation of the increment with localization using Eq (7) (i.e., $\boldsymbol{x}'_{m_x} = \boldsymbol{\rho}_{m_x \times m_y} \circ [\mathbf{b}_{m_x \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T] \boldsymbol{y}^{rl}_{m_y}$) needs $m_x \times m_y \times (n+2)$ multiplications and $m_x \times m_y \times n - m_x$ additions, while that without localization using Eq (6) (i.e., $\boldsymbol{x}'_{m_x} = [\mathbf{b}_{m_x \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T] \boldsymbol{y}^r_{m_y}$) takes only $(m_x + m_y) \times n$ multiplications and $(m_x + m_y - 1) \times n - m_x$ additions.

To provide an intuitive understanding of the huge cost of localization in the EnKF when assimilating multi-source observations, including satellite measurements, let us consider a typical realistic NWP configuration such that $m_x = 10^7$, $m_y = 10^5$ and $n = 30$. In this case, the calculation of the increment with localization takes about 3×10^{13} multiplications and about 3×10^{13} additions, much more expensive than without localization, which takes about 3×10^8 multiplications and 3×10^8 additions.

To reduce the huge cost of localization in the EnKF, two kinds of methods are frequently used, including batch processing [2] and serial processing [14]. In batch processing, the observations are organized into batches and each batch is assimilated simultaneously, while in serial processing, observations are assimilated sequentially, one by one. However, even if batch or serial processing is performed in the EnKF, the computational cost is still quite large in practice. For example, the EnKF with serial processing needs m_y sequential assimilations and, each time at least $m_x \times n$ multiplications are required according to the gain matrix formula (see Eq (6)). In total, at least $m_x \times m_y \times n$ multiplications are needed for the serial implementation of the EnKF. This is still a huge cost. As in the previous example, let m_x, m_y and n be 10^7, 10^5 and 30, respectively; the EnKF will perform at least 3×10^{13} multiplications. Obviously, the cost reduction by serial processing is not significant. The cost of batch processing can be estimated in the same way, and is just as large as the serial processing. The obvious advantage of these two methods is easier to obtain the solutions to $\boldsymbol{y}^{rl}_{m_y} = (\boldsymbol{\rho}_{m_y \times m_y} \circ \mathbf{P}^{\boldsymbol{y}}_{m_y \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T + \mathbf{R}_{m_y \times m_y})^{-1} \boldsymbol{y}'^{obs}_{m_y}$, which is much less costly than that of $\boldsymbol{x}'_{m_x} = \boldsymbol{\rho}_{m_x \times m_y} \circ [\mathbf{b}_{m_x \times n}(\mathbf{P}^{\boldsymbol{y}}_{m_y \times n})^T] \boldsymbol{y}^{rl}_{m_y}$.

In this paper, we proceed as follows: first a more detail descriptions of the new scheme are given. Then, we compare its filtering performance with the standard scheme. We also conduct preliminary tests using the new approach in the EnKF. Finally, we discuss our methods, the scope and limitations of this study, and some of the possible extension.

Materials and methods

Methodology

The basic idea of covariance localization is to limit the number of observations that can affect the analysis at a particular gridpoint. A simple technique for this is through observation selection, since the analysis is affected by observations within a cutoff radius [2]. Another way to implement covariance localization is to apply a Schür product between the forecast error covariance matrix and a correlation matrix [13], either in model space (Eq (3)) or in observation space (Eq (4)). For example, the element $\rho_{i,j}$ of the correlation matrix $\boldsymbol{\rho}_{m_x \times m_y}$ can be written as

$$\rho_{i,j} = C_0\left(\frac{d^h_{i,j}}{d^h_0}\right) \cdot C_0\left(\frac{d^v_{i,j}}{d^v_0}\right), \left(i = 1, 2, \ldots, m_x; j = 1, 2, \ldots, m_y\right), \tag{8}$$

where d^h_0 and d^v_0 are the prescribed horizontal and vertical filtering radii, respectively; $d^h_{i,j}$ and $d^v_{i,j}$ are the horizontal and vertical distances between the i-th control variable and the j-th observation, respectively. C_0 in (8) is the GC correlation function:

$$C_0(r) = C_0\left(L_i, L_j\right) = \begin{cases} -\frac{1}{4}r^5 + \frac{1}{2}r^4 + \frac{5}{8}r^3 - \frac{5}{3}r^2 + 1, & 0 \le r \le 1 \\ \frac{1}{12}r^5 - \frac{1}{2}r^4 + \frac{5}{8}r^3 + \frac{5}{3}r^2 - 5r + 4 - \frac{2}{3}r^{-1}, & 1 < r \le 2 \\ 0, \quad 2 < r \end{cases} \tag{9}$$

where $= \frac{d_{i,j}}{d_0}$, L_i and L_j are the spatial coordinates of the i-th control variable and the j-th observation, respectively. From Eqs (6) and (7), it is clear that the increase in computational cost due to the localization is mainly caused by the Schür product between $\mathbf{b}_{m_x \times n}(\mathbf{P}^y_{m_y \times n})^T$ and the correlation matrix $\boldsymbol{\rho}_{m_x \times m_y}$, leading to the change of $\mathbf{b}_{m_x \times n}(\mathbf{p}^y_{m_y \times n})^T$ from a separable form to an inseparable form. This expensive calculation can fully be avoided and can be reduced to a time-saving product between $\mathbf{b}_{m_x \times n}$ and an n-dimensional vector if the Schür product is not performed.

If the localization matrix $\boldsymbol{\rho}_{m_x \times m_y}$ can possibly be decomposed into a product of two vectors:

$$\boldsymbol{\rho}_{m_x \times m_y} = \boldsymbol{\rho}^x_{m_x}(\boldsymbol{\rho}^y_{m_y})^T, \tag{10}$$

the aforementioned Schür product may become separable:

$$\boldsymbol{\rho}_{m_x \times m_y} \circ [\mathbf{b}_{m_x \times n}(\mathbf{P}^y_{m_y \times n})^T] = \tilde{\mathbf{b}}_{m_x \times n}(\tilde{\mathbf{P}}^y_{m_y \times n})^T, \tag{11}$$

where $\boldsymbol{\rho}^x_{m_x}$ and $\boldsymbol{\rho}^y_{m_y}$ are m_x- and m_y-dimension vectors, respectively, and $\tilde{\mathbf{b}}_{m_x \times n}$ and $\tilde{\mathbf{P}}^y_{m_y \times n}$ are:

$$\begin{cases} \tilde{\mathbf{b}}_{m_x \times n} = \frac{1}{\sqrt{n-1}}(\boldsymbol{\rho}^x_{m_x} \circ (\mathbf{x}_1 - \bar{\mathbf{x}}), \boldsymbol{\rho}^x_{m_x} \circ (\mathbf{x}_2 - \bar{\mathbf{x}}), \ldots, \boldsymbol{\rho}^x_{m_x} \circ (\mathbf{x}_n - \bar{\mathbf{x}})) \\ \tilde{\mathbf{p}}^y_{m_y \times n} = \frac{1}{\sqrt{n-1}}(\boldsymbol{\rho}^y_{m_y} \circ (\mathbf{y}_1 - \bar{\mathbf{y}}), \boldsymbol{\rho}^y_{m_y} \circ (\mathbf{y}_2 - \bar{\mathbf{y}}), \ldots, \boldsymbol{\rho}^y_{m_y} \circ (\mathbf{y}_n - \bar{\mathbf{y}})) \end{cases}. \tag{12}$$

In this way, the high computational cost resulting from the localization can be greatly reduced.

However, this localization matrix cannot be expressed in the form of Eq (10), because it is impossible to decompose the correlation function into the following form:

$$C_0(r) = C_0(L_i, L_j) = C_0^x(L_i) \cdot C_0^y(L_j), \tag{13}$$

according to its definition by Eq (9). How to decompose the correlation function becomes the key point to reduce high computational cost for localization. Liu et al. [21] used the empirical orthogonal function (EOF) to decompose the correlation function on a low-resolution grid, and then interpolated the chosen dominant modes to the high-resolution grid. This is one of the earliest studies to expand the GC localization function. It is an efficient method that avoids the high cost of conducting the EOF on the high-resolution model grid directly, but it inevitably results in a reduction of accuracy in calculating the correlation function, due to the low precision of the leading modes decomposed on the low-resolution grid and the interpolation from the low-resolution grid to the high-resolution grid. Buehner et al. [31–32], Bishop et al. [33] and Kuhl et al. [34] adopted scaled spherical harmonics to decompose the correlation function. They provided an analytical and continuous expansion of the GC localization function. However, it is difficult to apply these methods to regional assimilations, because the spherical harmonics that requires homogeneous or periodic boundary conditions is more suitable for a spherical domain. Actually, due to the same reason, regional models rarely use the spherical harmonics for discretization of their dynamical cores. To avoid the aforesaid problems, this study tries to find a group of basis functions to expand the correlation function:

$$C_0(r) = C_0(\mathbf{L}_i, \mathbf{L}_j) = \sum_{k=1}^{K_c} \beta_k e_k(\mathbf{L}_i) e_k(\mathbf{L}_j), \tag{14}$$

so that the expansion is applicable for assimilations in both spherical and rectangular domains. The basis function $e_k(\mathbf{L})$ in (14) is analytical and subject to orthogonality as follows:

$$\int_\Omega w(\mathbf{L}) e_k(\mathbf{L}) e_l(\mathbf{L}) ds = \begin{cases} 1 & \text{if} k = l \\ 0 & \text{if} k \neq l \end{cases}, \tag{15}$$

where $w(x)$ is a weighting function and Ω is the domain of the model. The coefficient β_k, which is the eigenvalue (or variance) of the k-th basis function, can be calculated directly according to the above orthogonality (Eq (15)):

$$\beta_k = \int_\Omega \int_\Omega w(\mathbf{L}_1) w(\mathbf{L}_2) C_0(\mathbf{L}_1, \mathbf{L}_2) e_k(\mathbf{L}_1) e_k(\mathbf{L}_2) ds_1 ds_2. \tag{16}$$

K_c in (14) is the number of basis functions, which is either infinite when β_k is calculated based on Eq (16), or a finite positive integer depending on the given resolution to discretely calculate the coefficient. If the orthogonal basis functions are just the intrinsic modes of the correlation function, a finite number of leading modes can be chosen to express the correlation function approximately as follows:

$$C_0(r) = C_0(\mathbf{L}_i, \mathbf{L}_j) \approx C_{K_0}(\mathbf{L}_i, \mathbf{L}_j) = \sum_{k=1}^{K_0} \beta_k e_k(\mathbf{L}_i) e_k(\mathbf{L}_j), \tag{17}$$

where K_0 is the number of selected leading modes, which should be much smaller than K_c. K_0 can be determined according to a given criterion for the contribution of accumulated variance of the chosen leading modes to the total variance: $\sum_{k=1}^{K_0} \beta_k \Big/ \sum_{k=1}^{K_c} \beta_k$ (say, 95% or more). In this

way, the localization matrix $\boldsymbol{\rho}_{m_x \times m_y}$ can be simplified to the following form:

$$\boldsymbol{\rho}_{m_x \times m_y} \approx \sum_{k=1}^{K_0} \boldsymbol{\rho}_{m_x}^{(x,k)} (\boldsymbol{\rho}_{m_y}^{(y,k)})^T. \tag{18}$$

In Eq (18), $\boldsymbol{\rho}_{m_x}^{(x,k)}$ and $\boldsymbol{\rho}_{m_y}^{(y,k)}$ are m_x-dimension and m_y-dimension vectors, respectively. Now, the localization can be reduced into

$$\begin{cases} \boldsymbol{\rho}_{m_x \times m_y} \circ [\mathbf{b}_{m_x \times n} (\mathbf{P}^{y}_{m_y \times n})^T] \approx \sum_{k=1}^{K_0} \mathbf{b}^{(k)}_{m_x \times n} (\mathbf{P}^{(k)}_{m_y \times n})^T \\ \boldsymbol{\rho}_{m_y \times m_y} \circ [\mathbf{P}^{y}_{m_y \times n} (\mathbf{P}^{y}_{m_y \times n})^T] \approx \sum_{k=1}^{K_0} \mathbf{P}^{(k)}_{m_y \times n} (\mathbf{P}^{(k)}_{m_y \times n})^T \end{cases}, \tag{19}$$

where

$$\begin{cases} \mathbf{b}^{(k)}_{m_x \times n} = \dfrac{1}{\sqrt{n-1}} (\boldsymbol{\rho}_{m_x}^{(x,k)} \circ (\mathbf{x}_1 - \bar{\mathbf{x}}), \boldsymbol{\rho}_{m_x}^{(x,k)} \circ (\mathbf{x}_2 - \bar{\mathbf{x}}), \ldots, \boldsymbol{\rho}_{m_x}^{(x,k)} \circ (\mathbf{x}_n - \bar{\mathbf{x}})) \\ \mathbf{p}^{(k)}_{m_y \times n} = \dfrac{1}{\sqrt{n-1}} (\boldsymbol{\rho}_{m_y}^{(y,k)} \circ (\mathbf{y}_1 - \bar{\mathbf{y}}), \boldsymbol{\rho}_{m_y}^{(y,k)} \circ (\mathbf{y}_2 - \bar{\mathbf{y}}), \ldots, \boldsymbol{\rho}_{m_y}^{(y,k)} \circ (\mathbf{y}_n - \bar{\mathbf{y}})) \end{cases}. \tag{20}$$

The new localization scheme mainly needs $(m_x + m_y) \times n \times K_0 \times 2$ multiplications and $((m_x + m_y - 1) \times n - m_x) \times K_0$ additions. Under the same resolutions as mentioned in the introduction ($m_x = 10^7$, $m_y = 10^5$, and $n = 30$), and if $K_0 = 20$, the multiplication and addition calculations are performed about 1.2×10^{10} and 6×10^9 times, respectively, in the new scheme. Its cost is roughly 2500 times lower than usual. In the next subsection, we will give the definition of the basis functions, and then investigate the precision of the expansion in the right side of Eq (17) in 1D and 2D cases.

Basis functions

As discussed above, expanding the correlation function by means of a group of basis functions can greatly reduce the cost of the localization. Therefore, the construction of basis functions is the first step for the expansion. Actually, the eigenvectors of the correlation function on the discrete grid with a prescribed resolution can be used to determine the features of the basis functions and, ultimately, to construct them analytically. For this purpose, the eigenvectors of the discrete correlation function are investigated first in the following. For convenience of discussion, investigations are conducted in the 1D case under periodic and non-periodic boundaries, respectively.

1D case under non-periodic boundary condition. Suppose the domain of definition is $[a, b]$, of which the length is $l_0 = b - a$. Uniformly partition the interval $[a, b]$ using m grids whose locations are $x_i = a+(i-1)\times dx$, where $dx = l_0/(m-1)$; $i = 1,2,\ldots, m$. For any x_i and x_j in $[a, b]$, their distance is defined as $d_{i,j} = |x_i - x_j|$. If the filtering radius is d_0, the non-dimensional distance can be expressed as $r_{i,j} = d_{i,j}/d_0 = |x_i - x_j|/d_0$. In this way, the value of the correlation function between the grids can be calculated according to Eq (2): $c_{i,j} = C_0(r_{i,j})$ ($i = 1,2,\ldots, m$; $j = 1,2,\ldots, m$), which, as the elements, forms the localization matrix $\boldsymbol{\rho}^{non-periodic}_{m\times m}$, a sparse banded matrix. Because this matrix is symmetric, it has m real non-negative eigenvalues $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_m \geq 0$ and corresponding unit

orthogonal eigenvectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$, so that

$$\boldsymbol{\rho}_{m\times m}^{non-periodic} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1m} \\ c_{21} & c_{22} & \ldots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \ldots & c_{mm} \end{bmatrix} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m) \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_m \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_m^T \end{bmatrix}$$
$$= \sum_{k=1}^{m} \sigma_k \mathbf{s}_k \mathbf{s}_k^T. \quad (21)$$

When m is very large, K_0 leading eigenvectors with the largest eigenvalues can be chosen to approximately expand the localization matrix:

$$\boldsymbol{\rho}_{m\times m}^{non-periodic} \approx \sum_{k=1}^{K_0} \sigma_k \mathbf{s}_k \mathbf{s}_k^T. \quad (22)$$

This means that the eigenvectors of the localization matrix can be the best choice for the basis functions in the discrete case. Therefore, we are interested in what analytical forms they have.

Set $a = -5.0$, $b = 5.0$, $d_0 = 1.0$, and $m = 101$. The eigenvectors can be easily calculated under this resolution, and their spatial distributions can also be depicted. For example, the spatial distributions of the first three eigenvectors (black solid lines) are shown in Fig 1. They are very close to the sine waves (see red dotted lines) $\sin\frac{k\pi}{l}(x_i - \tilde{a})$ ($k = 1, 2, 3$; $i = 1, 2, \ldots, m$) that are incomplete in the domain of definition and are defined on an extended domain $[\tilde{a}, \tilde{b}]$, where $\tilde{a} = a - \frac{l-l_0}{2}$, $\tilde{b} = b + \frac{l-l_0}{2}$, and $l > l_0$, because the values at the beginning and ending points of the interval $[a, b]$ are not zero. Furthermore, when the resolution is increased, the wave shapes of the eigenvectors change very little. Fig 2A shows the second eigenvectors as an example with different resolutions of $m = 101$ (black solid line), $m = 401$ (brown dashed line) and $m = 801$ (blue dotted line), respectively, which barely differ from each other. Their differences between two adjacent resolutions are much smaller than the eigenvectors themselves in terms of amplitude (Fig 2B). In particular, these differences become smaller as resolution increases (Fig 2B). This suggests that sine-function-based eigenvectors are insensitive to grid resolution. On the other hand, the relative change of the extended boundary, defined as $\varepsilon = \frac{l-l_0}{l_0}$, is in a small range [0.066, 0.075] when the resolution varies from $m = 101$ to $m = 801$. It indicates that the analytical forms of the eigenvectors can be sine functions with different frequencies approximately so that they can be used as the basis functions:

$$e_k(x) = \sin\frac{k\pi}{l}(x - \tilde{a}) \ (l = \tilde{b} - \tilde{a} = (1 + \varepsilon)l_0;\ k = 1, 2, \ldots, K_0). \quad (23)$$

These functions are orthogonal on the extended domain of definition $[\tilde{a}, \tilde{b}]$:

$$\int_{\tilde{a}}^{\tilde{b}} w(x)\sin\frac{k_1\pi}{l}(x - \tilde{a})\sin\frac{k_2\pi}{l}(x - \tilde{a})dx = \begin{cases} 1, & \text{if} k_1 = k_2 \\ 0, & \text{if} k_1 \neq k_2 \end{cases}, \quad (24)$$

where $w(x) = \frac{2}{l}$. Using the above sine functions, the correlation function can be approximately

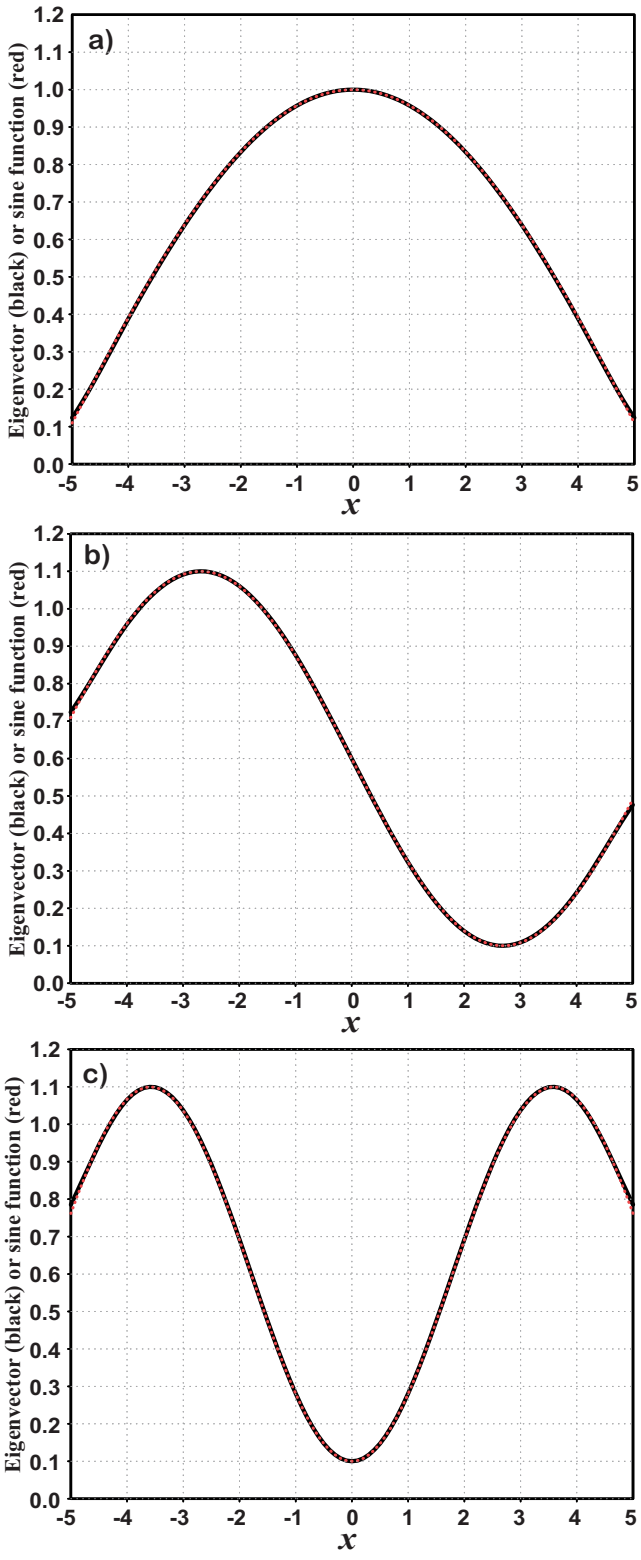

Fig 1. Leading eigenvectors of the correlation function (black solid) and sine function $\sin\frac{k\pi}{l}(x-\tilde{a})$ (red dotted), including a) the first eigenvector and sine function with $k = 1$, b) the second eigenvector and sine function with $k = 2$, and c) the third eigenvector and sine function with $k = 3$. Note that the two curves overlap in each panel.

https://doi.org/10.1371/journal.pone.0191088.g001

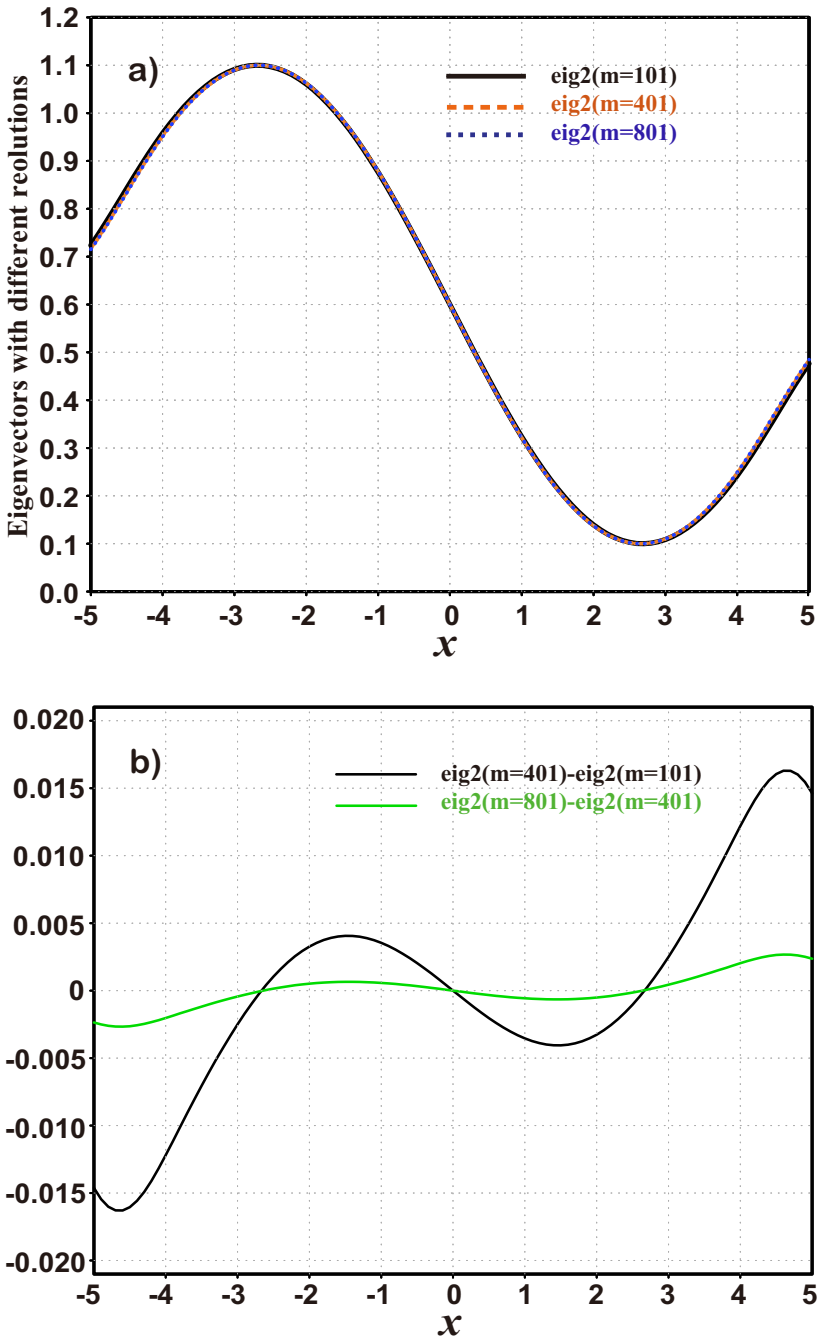

Fig 2. a) The second eigenvectors at different spatial resolutions including m = 101 (black solid), m = 401 (orange yellow dashed) and m = 801 (blue dotted); and b) the differences of eigenvectors between the resolutions of m = 101 and m = 401 (black line) and between the resolutions of m = 401 and m = 801.

https://doi.org/10.1371/journal.pone.0191088.g002

expressed by a truncated expansion:

$$C_0(r) = C_0(x_1, x_2) \approx C_{K_0}(x_1, x_2) = \sum_{k=1}^{K_0} \beta_k \sin\frac{k\pi}{l}(x_1 - \tilde{a}) \sin\frac{k\pi}{l}(x_2 - \tilde{a}), \quad (25)$$

where

$$\beta_k = \frac{4}{l^2}\int_{\tilde{a}}^{\tilde{b}}\int_{\tilde{a}}^{\tilde{b}} C_0(x_1, x_2)\sin\frac{k\pi}{l}(x_1 - \tilde{a})\sin\frac{k\pi}{l}(x_2 - \tilde{a})dx_1 dx_2. \quad (26)$$

1D case under periodic boundary condition. In this case, the domain of definition is supposed to be a zonal circle at any a latitude θ, i.e., the longitude λ varies from $a = 0$ to $b = 2\pi$. Uniformly partition the interval $[a, b]$ using m grids whose locations are $\lambda_i = (i-1)\times d\lambda$, where $d\lambda = 2\pi/(m-1)$; $i = 1,2,\ldots, m$. For any λ_i and λ_j on the zonal circle, their distance can be defined as their arc length: $d_{i,j} = R_0\cos\theta\times\min(|\lambda_i-\lambda_j|,2\pi-|\lambda_i-\lambda_j|)$ because of the periodic boundary, where R_0 is the radius of the Earth. When the filtering radius of longitude is λ_0, the geometrical filtering radius is then $d_0 = R_0\cos\theta\times\lambda_0$. Consequently, the non-dimensional distance between λ_i and λ_j is $r_{i,j} = d_{i,j}/d_0 = \min(|\lambda_i-\lambda_j|,2\pi-|\lambda_i-\lambda_j|)/\lambda_0$, with which the localization matrix $\boldsymbol{\rho}_{m\times m}^{periodic}$ calculated according to Eq (9) is still a symmetric matrix, but not a banded matrix. Similarly, it has m real non-negative eigenvalues $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq 0$ and corresponding unit orthogonal eigenvectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$, so that

$$\begin{aligned}\boldsymbol{\rho}_{m\times m}^{periodic} &= \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1m} \\ c_{21} & c_{22} & \ldots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \ldots & c_{mm} \end{bmatrix} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m)\begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_m \end{bmatrix}\begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_m^T \end{bmatrix} \\ &= \sum_{k=1}^{m}\sigma_k\mathbf{s}_k\mathbf{s}_k^T \end{aligned} \quad (27)$$

As decomposed above, the spatial distributions of the eigenvectors can also approximately be expressed as sine functions with different frequencies and phases. They can be defined with a periodic domain $[a, b]$ as:

$$e_k(\lambda) = \sin\left(\frac{k\pi}{l}(\lambda - a) + \omega_k\right)\ (l = b - a;\ k = 1, 2, \ldots, K_0). \quad (28)$$

These functions are orthogonal in the domain of definition $[a,b]$:

$$\int_a^b w(\lambda)\sin\left(\frac{k_1\pi}{l}(\lambda - a) + \omega_{k_1}\right)\sin\left(\frac{k_2\pi}{l}(\lambda - a) + \omega_{k_2}\right)d\lambda = \begin{cases} 1, & \text{if} k_1 = k_2 \\ 0, & \text{if} k_1 \neq k_2 \end{cases}, \quad (29)$$

where $w(\lambda) = \frac{2}{l}$. Using the above sine functions, the correlation function can be approximately expressed by a truncated expansion:

$$C_0(r) = C_0(\lambda_1, \lambda_2) \approx C_{K_0}(\lambda_1, \lambda_2) = \sum_{k=1}^{K_0}\beta_k\sin\left(\frac{k\pi}{l}(\lambda_1 - a) + \omega_k\right)\sin\left(\frac{k\pi}{l}(\lambda_2 - a) + \omega_k\right), \quad (30)$$

where

$$\begin{cases} \omega_k = \tan^{-1}\left(\dfrac{\int_a^b\int_a^b C_0(\lambda_1, \lambda_2)\ \sin\dfrac{k\pi}{l}(\lambda_1 - a)\ \cos\dfrac{k\pi}{l}(\lambda_2 - a)\ d\lambda_1 d\lambda_2}{\int_a^b\int_a^b C_0(\lambda_1, \lambda_2)\ \sin\dfrac{k\pi}{l}(\lambda_1 - a)\ \sin\dfrac{k\pi}{l}(\lambda_2 - a)\ d\lambda_1 d\lambda_2}\right) \\ \beta_k = \dfrac{4}{l^2}\int_a^b\int_a^b C_0(\lambda_1, \lambda_2)\ \sin\left(\dfrac{k\pi}{l}(\lambda_1 - a) + \omega_k\right)\ \sin\left(\dfrac{k\pi}{l}(\lambda_2 - a) + \omega_k\right)\ d\lambda_1 d\lambda_2 \end{cases}. \quad (31)$$

Distance functions

The distance function or the non-dimensional distance function is critical for formation of the localization matrices $\boldsymbol{\rho}_{m\times m}^{non-periodic}$ and $\boldsymbol{\rho}_{m\times m}^{periodic}$ in 1D cases based on Gaspari and Cohn [28]. If a 1D model has equal spacing grids, as supposed in the aforesaid 1D cases under periodic and non-periodic boundary conditions, the non-dimensional distances can be expressed as functions with respect to the grid number (i). For example, in the 1D periodic case, the non-dimensional distance between two model grid-points can be formulated as $r_{i,j} = d_{i,j}/d_0 = \min(|i-j|, m-1-|i-j|)/i_0$, according to the expressions $\lambda_i = (i-1)\times d\lambda$, $\lambda_j = (j-1)\times d\lambda$ and $2\pi = (m-1)\times d\lambda$, where $i_0 = d_0/d\lambda$. Similarly, the non-dimensional distance between two model grid-points in the 1D non-periodic case can be expressed as $r_{i,j} = d_{i,j}/d_0 = |i-j|/i_0$, where $i_0 = d_0/dx$. For the non-dimensional distances between a model grid-point and an observation location, the subscript j may be a real number: $j = 1 + \lambda_j/d\lambda$, where λ_j is the observation location.

In 2D cases, the non-dimensional distance between two model grid-points can also be expressed as functions of grid number (i, j). If the domain is rectangular with $m\times n$ discrete grids, i.e., $[a, b; c, d]$ with the grid-sizes $dx = (b-a)/(m-1)$ and $dy = (d-c)/(n-1)$, respectively, the distance between any two discrete points $A(x_{i_A}, y_{j_A})$ and $B(x_{i_B}, y_{j_B})$ in this domain can be defined as $d_{A,B} = \sqrt{(x_{i_A} - x_{i_B})^2 + (y_{j_A} - y_{j_B})^2} = \sqrt{(i_A - i_B)^2 dx^2 + (j_A - j_B)^2 dy^2}$, where $x_i = a+(i-1)\times dx$ and $y_i = c+(j-1)\times dy$. The corresponding non-dimensional distance is then expressed as $r_{A,B} = d_{A,B}/d_0 = \sqrt{r_x^2 + r_y^2}$, where d_0 is the filtering radius, $r_x = |i_A - i_B|/i_0$, $r_y = |j_A - j_B|/j_0$, $i_0 = d_0/dx$ and $j_0 = d_0/dy$. Because the correlation function $C_0(r)$ has a close relationship with the exponential function [35]:

$$C_0(r_{A,B}) \sim e^{-\alpha r_{A,B}^2} = (e^{-\alpha r_x^2})(e^{-\alpha r_y^2}), \tag{32}$$

where the constant $\alpha > 0$, the correlation function is approximately separable in 2D cases:

$$C_0(r_{A,B}) \approx C_0(r_x) \cdot C_0(r_y). \tag{33}$$

Therefore, we assume that the 2D expansion $\tilde{C}_{K_0}(r)$ is approximately separable:

$$\tilde{C}_{K_0}(r_{A,B}) \approx \tilde{C}_{K_0}(r_x) \cdot \tilde{C}_{K_0}(r_y). \tag{34}$$

It suggests that a 2D correlation function can be calculated using two 1D correlation function, which greatly reduces its complexity in calculations. If (i, j) is used to express an observation location, i and j may not be integer numbers, as in the 1D case.

If the 2D domain is the spherical surface with longitude-latitude coordinates: $(\lambda, \theta)\in[0, 2\pi; -\pi/2, \pi/2]$, which is widely used in global atmospheric models, the exact distance between two discrete points $A(\lambda_{i_A}, \theta_{j_A})$ and $B(\lambda_{i_B}, \theta_{j_B})$ in this domain is defined as $R_0\cos^{-1}(\sin\theta_{j_A}\sin\theta_{j_B} + \cos\theta_{j_A}\cos\theta_{j_B}\cos(\lambda_{i_A} - \lambda_{i_A}))$. However, this formula of distance may lead to inseparability in calculation of the corresponding correlation function. Due to this reason, the distance function here is approximately defined as the hypotenuse of the curved-edge right triangle consisting of the points A, B and O, where the point O can be $O_A(\lambda_{i_B}, \theta_{j_A})$ or $O_B(\lambda_{i_A}, \theta_{j_B})$. It means two right triangles ΔBAO_A and ΔABO_B share the same hypotenuse. These two triangles have the same meridional leg length but different lengths of zonal leg ($\overline{AO_A}$ and $\overline{O_BB}$), which are $d_\lambda^A = R_0\cos\theta_{j_A} \times \min(|\lambda_{i_A} - \lambda_{i_B}|, 2\pi - |\lambda_{i_A} - \lambda_{i_B}|)$ and $d_\lambda^B = R_0\cos\theta_{j_B} \times \min(|\lambda_{i_A} - \lambda_{i_B}|, 2\pi - |\lambda_{i_A} - \lambda_{i_B}|)$, respectively. The value of the exact distance $d_{A,B}$ is between the hypotenuse lengths of ΔBAO_A and ΔABO_B that are respectively $d_{A,B}^A = \sqrt{(d_\lambda^A)^2 + d_\theta^2}$ and $d_{A,B}^B = \sqrt{(d_\lambda^B)^2 + d_\theta^2}$

(where $d_\theta = R_0|\theta_{j_A} - \theta_{j_B}|$), i.e., $\min(d^A_{A,B}, d^B_{A,B}) < d_{A,B} < \max(d^A_{A,B}, d^B_{A,B})$. Because the difference between $d^A_{A,B}$ and $d^B_{A,B}$ is completely due to the difference between d^A_λ and d^B_λ resulted from their different latitudes θ_{j_A} and θ_{j_B}, the zonal arc length at the middle of two latitudes $\theta_M = (\theta_{j_A} + \theta_{j_B})/2$, which is $d_\lambda = R_0\cos\theta_M \times \min(|\lambda_{i_A} - \lambda_{i_B}|, 2\pi - |\lambda_{i_A} - \lambda_{i_B}|)$, is used to approximately define the distance between A and B: $d_{A,B} \approx \sqrt{d^2_\lambda + d^2_\theta}$. The corresponding non-dimensional distance can similarly be expressed using grid numbers (i, j): $r_{A,B} = d_{A,B}/d_0 \approx \sqrt{r^2_\lambda + r^2_\theta}$, where $r_\lambda = \min(|i-j|, m-1-|i-j|)/i_\theta$, $r_\theta = |j_A - j_B|/j_0$, $i_\theta = d_0/(R_0 d\lambda \cos\theta_M)$ and $j_0 = d_0/(R_0 d\theta)$. In this way, the correlation function for localization in the spherical domain can then be computed using 1D correlation functions according to Eq (34).

To provide an intuitive evaluation on how much the approximation of distance is, we consider a spherical domain with grid-sizes of 4.5˚ × 4.5˚, which is used by the spherical barotropic model in section 3.2. One location is selected at the equator, and the other, at a higher latitude. Table 1 gives the exact arc length between two points on a sphere (the arc length $\overline{AB}$) and the approximate value calculated by $\sqrt{d^2_\lambda + d^2_\theta}$. We can see that the longer the distance between two points A and B, the larger the error; and the error at the higher latitude is larger than that near the equator. For example, the error of the distance between point A (90˚N, 120˚E) and point B (72˚N, 138˚E) at high latitudes is about 24 km, so the relative error is no more than 2.4%. Compared with the filtering radius used to define non-dimensional distance (e.g., eight grids used by the spherical barotropic model in next section), the influence of such error is much smaller and negligible. The distance errors at lower latitudes are even smaller.

Preliminary evaluation

Given the analytical basis functions shown in Eq (25), numerical tests are conducted to evaluate the expansions of the correlation function with different truncations through comparison with the original one, in the 1D and 2D cases, respectively.

1D case. As defined in section 2.2, the non-dimensional distance between any $x \in [a, b]$ and a prescribed $x_0 = \in [a, b]$ is expressed as $r = |x - x_0|/d_0$, where $a = -5$, $b = 5$, and $d_0 = 1.0$. Setting $x = x_i$ (i = 1,2,. . ., m;m = 101) and $x_0 = x_{i_0}$ (i_0 can be any integer number on [1, m]; here, we select 49), the original correlation function $C_0(r_i)$ (black curves in Fig 3) and its expansion $C_{K_0}(x, x_0)$ (green curves in Fig 3) with different truncation numbers K_0 are then calculated on the discrete grids. It is found that the larger the truncation number K_0 is, the closer to the truth the expansion gets (Fig 3). We can clearly see some fluctuations along the true curve at the location where the correlation coefficients are very small when K_0 = 10 (Fig 3A). As the truncation number increases, these fluctuations become obviously weaker as K_0 = 15 (Fig 3B), and ultimately disappear when using K_0 = 20 (Fig 3C). This means that the first 20 modes form the dominant part of the localization function. In terms of variance contribution, the 20 leading modes account for more than 97% of all modes, no matter how high the resolution becomes (e.g., m = 1001, 10001. see Table 2). In other words, a large number of the remaining modes account for less than 3% of all modes. Table 3 shows that many of the modes have very

Table 1. The exact arc length ($\overline{AB}$) and the approximate value calculated using $\sqrt{d^2_\lambda + d^2_\theta}$ between two points A and B on a sphere.

A	B	The arc length $\overline{AB}$	Approximate value
(0˚N, 120˚E)	(9˚N, 129˚E)	1412.358	1413.101
(0˚N, 120˚E)	(18˚N, 138˚E)	2806.875	2813.190
(90˚N, 120˚E)	(81˚N, 129˚E)	1000.754	1003.830
(90˚N, 120˚E)	(72˚N, 138˚E)	2001.509	2025.851

https://doi.org/10.1371/journal.pone.0191088.t001

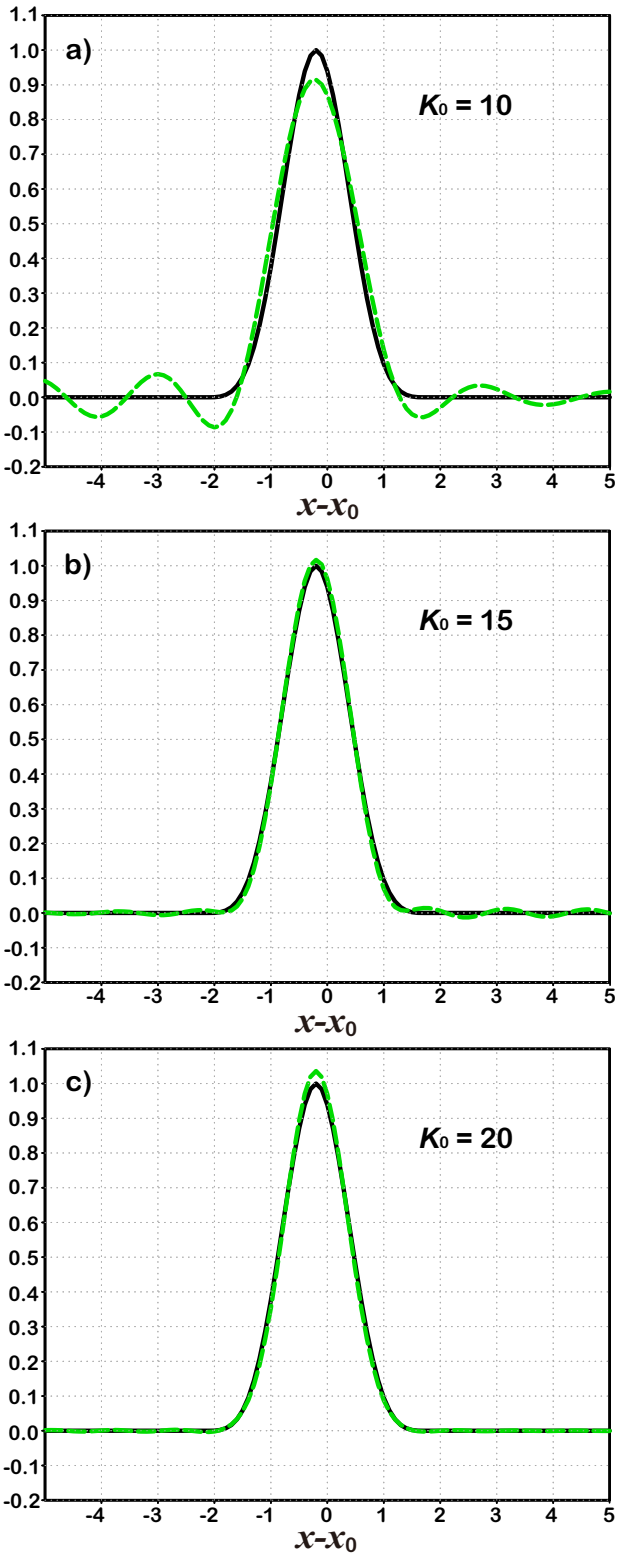

Fig 3. Comparisons between 1D filters presented by the correlation function (black solid) and the expansions (green dashed) with different truncations: K = 10 (a), 15 (b), and 20 (c).

https://doi.org/10.1371/journal.pone.0191088.g003

Table 2. Contributions of accumulated variances of leading modes to the total variance in the 1D case.

Number of leading modes used	Ratio of accumulated variance in the total variance		
	$m = 101$	$m = 1001$	$m = 10001$
1	12.34%	12.14%	12.13%
2	24.29%	28.95%	23.88%
3	35.61%	35.05%	35.02%
4	46.11%	45.41%	45.37%
5	55.62%	54.81%	54.77%
6	64.06%	63.17%	63.12%
7	71.37%	70.43%	70.37%
8	77.55%	76.58%	76.52%
9	82.64%	81.67%	81.61%
10	86.73%	85.78%	85.72%
11	89.94%	89.02%	88.95%
12	92.38%	91.50%	91.43%
13	94.19%	93.35%	93.27%
14	95.50%	94.69%	94.62%
15	96.43%	95.65%	95.57%
16	97.07%	96.32%	96.24%
17	97.51%	96.78%	96.71%
18	97.83%	97.11%	97.04%
19	98.06%	97.35%	97.28%
20	98.24%	97.54%	97.47%
Total variance	1.982459	1.998261	1.999826

https://doi.org/10.1371/journal.pone.0191088.t002

small eigenvalues, which make very small contributions to the correlation function when fitting observations. Therefore, with a reasonably small number of modes, the new localization will save computational time without sacrificing much accuracy.

Considering a 1D case under periodic boundary condition, with the same experiment setup, but set $x_0 = x_{i_0}$ and select i_0 to be 90. Fig 4 shows the original correlation function $C_0(r_i)$ (black curves) and its expansion $C_{K_0}(x, x_0)$ (green curves) with different truncation number K_0. Consistent with our finding in the non-periodic case, increase of the truncation number leads to higher accuracy of the value calculated through our expansion.

2D case. In this case, the non-dimensional distance between two points (x, y) and (x_0, y_0) is defined as $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}/d_0$, where $x, x_0 \in [a, b]$ and $y, y_0 \in [e, f]$. Similar to the 1D case, we set $a = e = -5$, $b = f = 5$, and $d_0 = 1$, and uniformly partition $[a, b]$ and $[e, f]$ using m grids, where $m = 101$. The prescribed point (x_0, y_0) is set to be $x_0 = x_{i_0}$ and $y_0 = y_{j_0}$, and the numbers i_0 and j_0 are selected to be 51.

Fig 5 illustrates a comparison among the 2D filters, separately presented by the expansion $\tilde{C}_{K_0}(r)$ with different truncations calculated according to Eq (30) (e.g. Fig 5A, Fig 5B and Fig 5C) and the original correlation function defined by Eq (9). The conclusion is similar to that in the 1D case, i.e., the larger the truncation number K_0 gets, the closer to the original correlation function the expansion becomes.

Assimilation experiments

The above section demonstrates the consistency between the sin basis function and the GC correlation function. Here, we will check it further with some assimilation experiments. An

experiment is that assimilating all observations simultaneously in the EnKF with the new localization approach, another way to do, the same as that in some numerical forecast centers, is to assimilate observation serially, one at a time in the EnKF scheme with the GC correlation function. The assimilation experiments are preliminarily tested using observation system simulation experiments (OSSEs) in two models that have increasing complexity: a Lorenz-96 model [36] and a spherical barotropic shallow water model. The "true" state (or "truth") is defined by a long-term model run, and the corresponding "observations" are generated by adding uncorrelated random noises to the "truth."

Lorenz-96 40-variable model. This model has been widely used to test ensemble-based assimilation methods in a number of earlier studies [14,37]. It is based on the following set of differential equations:

$$\frac{dx_j}{dt} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F, \tag{35}$$

where $j = 1,2,\ldots, M$ is the spatial coordinate; the forcing parameter and the number of spatial elements are set to $F = 8$ and $M = 40$, respectively. The model solves Eq (35) using the fourth-order Runge–Kutta scheme with a time-step of 0.05, where the boundary conditions of Eq (30) are periodic: $x_{j+M} = x_j$ [38]. Simulations during a period of time after a long-term integration (e.g., 10^5 model time steps) of the model from an arbitrary initial condition are assumed to be the "truth". Observational data sets include observations of all model variables that are produced by adding uncorrelated random noises with the standard Gaussian distribution (with zero mean and variance of 4.0) to the truth at every step. In this case, the observation number

Table 3. Variances (or eigenvalues) of representative modes in the 1D case.

Mode number	Variance (or eigenvalue)		
	$m = 101$	$m = 1001$	$m = 10001$
1	2.445878E-01	2.425069E-01	2.425069E-01
2	2.368530E-01	2.349722E-01	2.349723E-01
3	2.244613E-01	2.228930E-01	2.228932E-01
4	2.081164E-01	2.069435E-01	2.069438E-01
5	1.887161E-01	1.879857E-01	1.879862E-01
6	1.672772E-01	1.669981E-01	1.669988E-01
7	1.448537E-01	1.449984E-01	1.449992E-01
8	1.224568E-01	1.229669E-01	1.229679E-01
9	1.009850E-01	1.017799E-01	1.017810E-01
10	8.116955E-02	8.215678E-02	8.215796E-02
11	6.354103E-02	6.462652E-02	6.462777E-02
12	4.841677E-02	4.951431E-02	4.951560E-02
13	3.590897E-02	3.694751E-02	3.694882E-02
14	2.595007E-02	2.687822E-02	2.687952E-02
15	1.833037E-02	1.911801E-02	1.911930E-02
16	1.274246E-02	1.337957E-02	1.338084E-02
17	8.826586E-03	9.319923E-03	9.321153E-03
18	6.212055E-03	6.580454E-03	6.581651E-03
19	4.551000E-03	4.820176E-03	4.821343E-03
20	3.542190E-03	3.739829E-03	3.740968E-03
100	9.163717E-06	8.302773E-05	8.402847E-05
1000	/	6.912985E-08	8.006075E-07
10000	/	/	2.463610E-10

https://doi.org/10.1371/journal.pone.0191088.t003

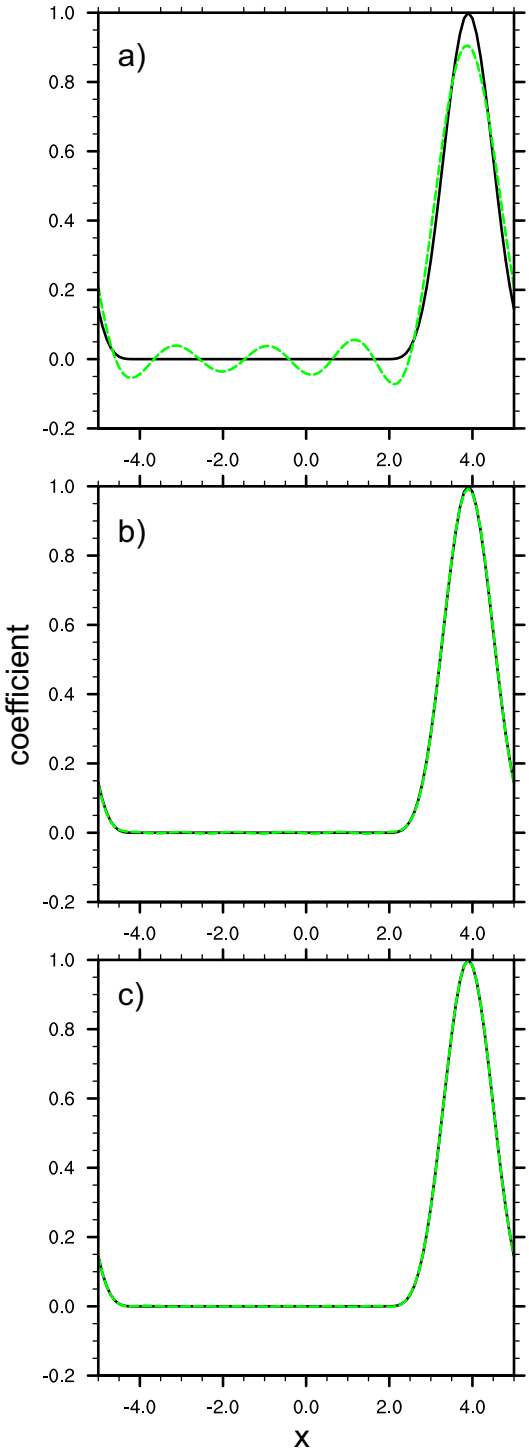

Fig 4. Same as Fig 3, except using a 1D periodic boundary.

https://doi.org/10.1371/journal.pone.0191088.g004

is 40, and no interpolation is needed. The observation error covariance matrix is diagonal. The EnKF is used to assimilate observations at each analysis time step in a cycle with a total of 800 time steps.

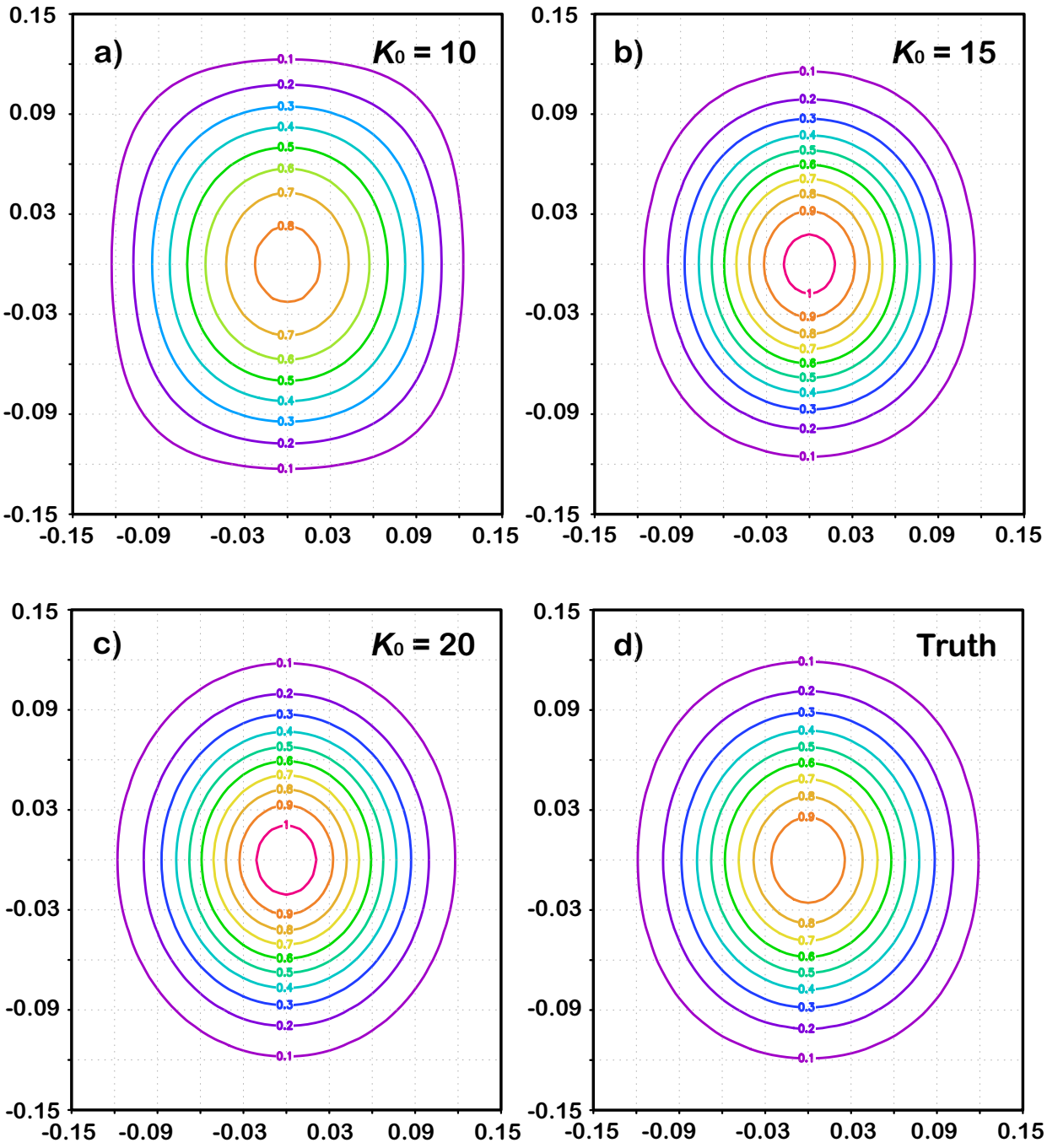

Fig 5. Comparisons between 2D filter presented by the GC correlation function (d) and the expansion with different truncations: $K = 10$ (a); $K = 15$ (b); and $K = 20$ (c).

https://doi.org/10.1371/journal.pone.0191088.g005

We conduct three assimilation experiments: one using 500 members without any localization (named "EXP-1"), and the other two using 20 members with the new (named "EXP-2") and traditional (named "EXP-3) localization schemes, respectively. The localization radius sets eight grid spacing, and all experiments use the covariance inflation method of Zhang et al.[10]:

$$(x_i^{'})_{new} = \alpha(x_i^{'})^f + (1-\alpha)(x_i^{'})^a, \tag{36}$$

where α is the relaxation coefficient; $(x_i')^a$ and $(x_i')^f$ denote the analysis and the perturbation of the i-th ensemble, respectively; and $(x_i')_{new}$ is the final perturbation of the updated ensemble members used for the next assimilation–forecast cycle. In these experiments, $\alpha = 0.15$.

To illustrate the differences among the three experiments more clearly, Fig 6 shows the root-mean-square errors (RMSEs) of the analysis only during the first 100 time steps. The results indicate that the new localization (EXP-2, red line) performs very similarly to the traditional one (EXP-3, black line). The RMSEs of both experiments with localizations are also close to those of the large-size ensemble experiment without the localization (EXP-1, blue line). In terms of overall performances of the three experiments in 800-step assimilation cycles, the new localization (EXP-2) generates the smallest error, of which the time-average RMSE over the 800-step cycles is 0.5558644. The time-averaged RMSEs of the other two experiments are 0.5934035 (EXP-1) and 0.5565553 (EXP-3). Unfortunately, the timesaving nature of the new localization is not obvious in the case of the simple model due to the low dimensions of control variables and observations. Both EXP-2 and EXP-3 used about 20 seconds, which is

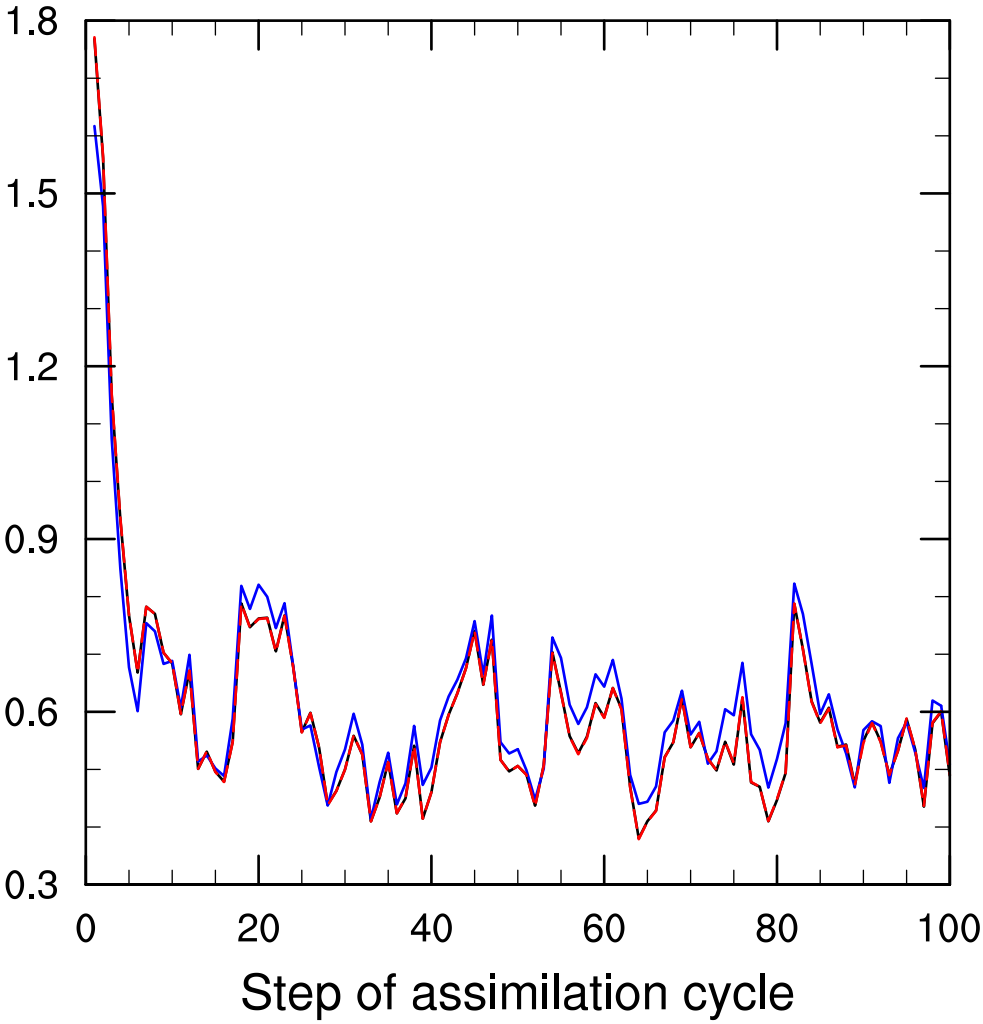

Fig 6. RMSEs of the analysis by the EnKF using 20 samples, respectively, with the new (red line) and traditional (black line) localization schemes and using 500 members without any localization (blue line) during the first 100 steps of the assimilation cycle.

https://doi.org/10.1371/journal.pone.0191088.g006

much more timesaving than the experiment with large ensemble size (EXP-1), which takes 428 seconds in the same computing environment. The significant timesaving characteristics of the new localization scheme become apparent in the following experiments with a more complex model.

Spherical barotropic shallow water model. To further compare the performances and computational costs of the new and traditional localizations, we use a spherical barotropic shallow water model to conduct two OSSEs. The model was established using a finite difference scheme with exact energy and mass conservations [39], to solve the following set of equations:

$$\begin{cases} \dfrac{\partial u}{\partial t}+\dfrac{u}{a\cos\theta}\dfrac{\partial u}{\partial \lambda}+\dfrac{v}{a}\dfrac{\partial u}{\partial \theta}+\dfrac{1}{a\cos\theta}\dfrac{\partial \varphi}{\partial \lambda}-fv=0 \\ \dfrac{\partial v}{\partial t}+\dfrac{u}{a\cos\theta}\dfrac{\partial v}{\partial \lambda}+\dfrac{v}{a}\dfrac{\partial v}{\partial \theta}+\dfrac{1}{a}\dfrac{\partial \varphi}{\partial \theta}+fu=0 \\ \dfrac{\partial \varphi}{\partial t}+\dfrac{1}{a\cos\theta}\left[\dfrac{\partial u\varphi}{\partial \lambda}+\dfrac{\partial v\cos\theta\varphi}{\partial \theta}\right]=0 \end{cases} . \tag{37}$$

Here, θ and λ are the latitude and longitude, respectively; u, v and φ denote zonal velocity, meridional velocity and geopotential height, respectively; a is the Earth's radius and f is the Coriolis coefficient.

The model has a horizontal resolution of $4.5^\circ \times 4.5^\circ$ (81×41grid points). The initial condition uses the four-wave Rossby–Haurwitz waves. A 20-day integration is conducted first, and the last 10-day integration is taken as the "truth" (i.e., nature run) after a 10-day spin-up. Synthetic observations of geopotential height are created every four gridpoints using the truth plus uncorrelated random noises with the standard Gaussian distribution (with zero mean and variance of 10000.0). In this way, there are 861 φ observations in all.

A common and easy way to implement the traditional localization in the EnKF is through serial processing [14], which assimilates the observations one by one in a cycle. This is considered to be more timesaving than the traditional localization, but with similar performance. Therefore, the serial implementation method is taken as the traditional localization here, and is compared with the new localization using the spherical barotropic model. Two OSSEs are designed for the comparison: one uses the traditional localization with serial implementation (called "ASSM_old"), and the other adopts the new localization with simultaneous implementation (named "ASSM_new"). The filtering radius in all experiments is eight grid spacing. Fig 7 compares the horizontal error distributions of geopotential height among the background (or first guess), and analyses from ASSM_old and ASSM_new. It shows that all analyses greatly reduce the phase errors of about 30° of longitude existed in the background. In addition, the analysis of ASSM_new at the higher latitude shows marked improvement, compared with ASSM_old. In terms of the RMSE, ASSM_new (1195.982) outperforms the traditional one (1437.468), while all analyses are much better than the background (2752.343). For the computational costs of the two localization schemes, the new localization uses only 25 seconds, far more timesaving than the traditional one, which needs 312 seconds in the same computing environment.

Discussions

In recent years, ensemble-based approaches have been widely used in various topics, e.g., data assimilation and solutions to conditional nonlinear optimal perturbation [40]. Because an ensemble is generally composed of far fewer members than both the number of observational data and the degrees of freedom of model variables, many spurious correlations between

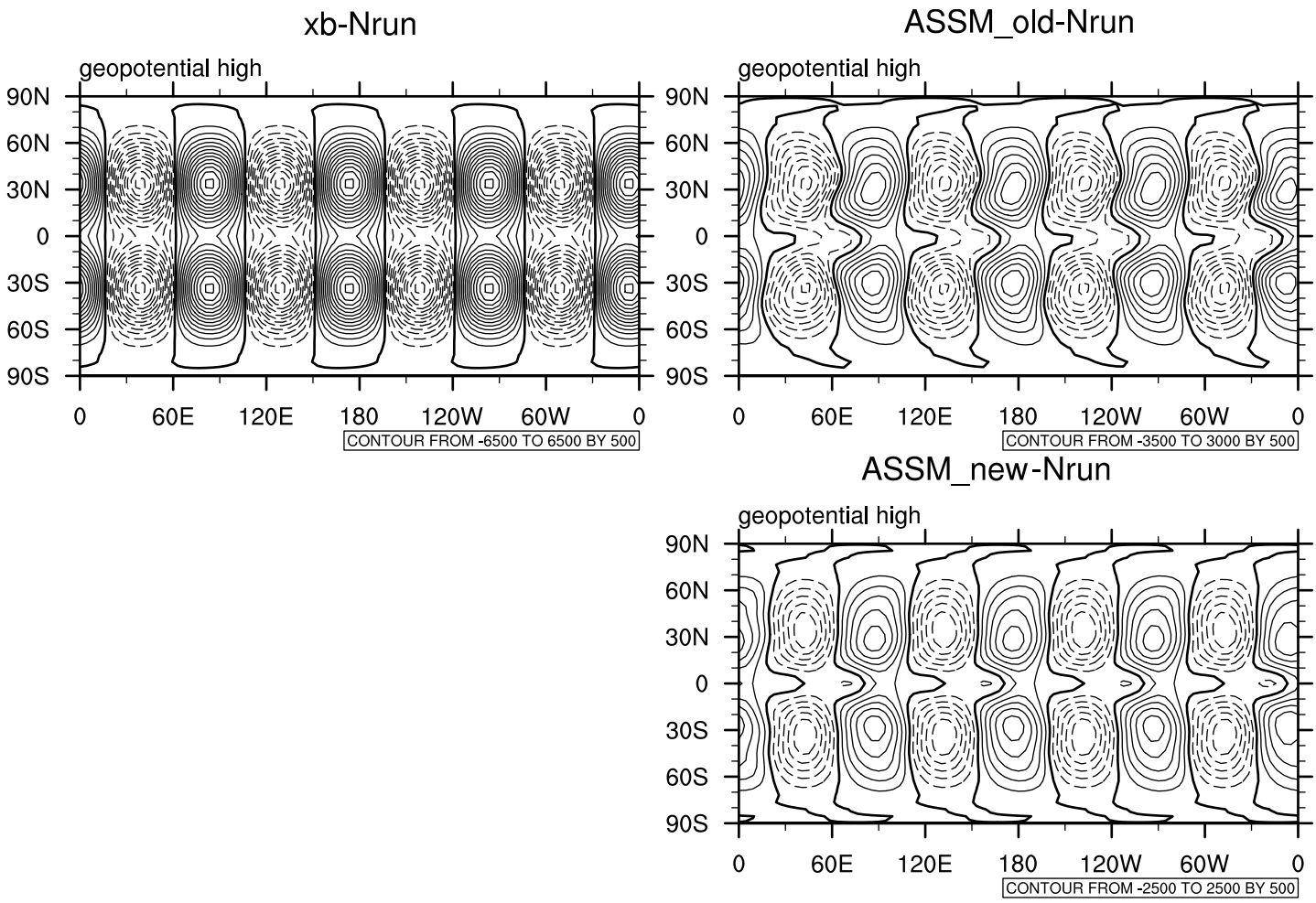

Fig 7. Comparison of the horizontal distributions of geopotential height errors among the background (or first guess), and analyses from ASSM_old and ASSM_new.

https://doi.org/10.1371/journal.pone.0191088.g007

different observation locations, between different model grids, or between observation locations and model grids, occurred. Schür product–based covariance localization has become a practical and powerful tool to make ensemble-based methods perform well even under small ensemble sizes [41]. However, a disadvantage of the traditional localization schemes is their large cost.

When observations assimilated are not many, there is little difference in computational cost between two localization schemes. However, with the large number of observations, even the serial implementation of EnKF, the computational cost is still increased dramatically. Then, if the localization uses a few basis functions to expand, it will be useful for improving work efficiency. This study is a preliminary attempt to develop and improve the localization approach within the EDA process. As the first and necessary step, the new scheme was preliminarily evaluated in its application to a simultaneous assimilation using idealized experiments. Further studies are required in three aspects. First, it is necessary to investigate its role in EDAs for real and complex forecast models. Second, an effort should be made to propose a new serial assimilation scheme in a way of using the leading modes one by one due to the orthogonality between these modes, similar to the way of assimilating the observations one by one in the serial processing of EnKF under the hypothesis of independence between these observations.

Third, it is worth exploring how to implement the adaptive localization approach [9,17], because it is now well understood that adaptive localization functions may be more appropriate, although the GC localization function has been widely used in EDA methods. It is anticipated that this work will be challenging due to the noticeable difference between the GC localization function, which is completely independent of ensemble samples, and the adaptive localization functions using complex corrections with respect to ensemble samples. This difference may lead to a great difficulty in expanding the adaptive localization functions using the sine functions as the basis functions, because various eigenvector families may be produced by different ensemble members in the adaptive localization.

Conclusions

In this paper, we proposed an economical approach to implement covariance localization. We attempted to use a group of basis functions to expand the correlation function, and found that the spatial distributions of the leading eigenvectors of the correlation function are very close to the sine waves that are defined in the domain of definition. We used the sine functions with different frequencies and phases approximately as the basis functions, so that the localization matrix can be decomposed into a series of products of two vectors, and then the Schür product is separable. In this way, the cost of localization can be greatly reduced.

Two numerical tests with different dimensions were conducted to evaluate the expansions of the correlation function. Both tests demonstrated that the larger the truncation number gets, the closer to the original correlation function the expansion becomes. When the truncation number reaches 20, the difference between the expansion and the truth is very small.

The scheme was then verified in an assimilation cycle with the Lorenz-96 model and a single assimilation experiment with a spherical barotropic shallow water model, using OSSEs of the EnKF. In general, when the ensemble size is much larger than the dimension of the model (e.g., 500 for a simple model like the Lorenz-96), the localization has no influence on the assimilation results and is thus not needed. However, if the ensemble size is smaller than the dimension of the model (say, 20), localization is necessary. The experiments conducted using the simple model suggested that ensemble assimilation using a smaller ensemble size with the new localization scheme could achieve a performance comparable to that with the traditional localization, and that applying a large ensemble size without any localization. The new localization even outperformed the traditional one with serial processing in the OSSEs using the spherical barotropic shallow water model. Moreover, the computational cost depends on the number of ensemble members, i.e., the larger the ensemble size gets, the higher the cost becomes. The new localization was shown to be far more timesaving than the serial implementation of the traditional localization in the single assimilation experiments using the spherical barotropic shallow water model, although the timesaving characteristics of the new localization was insignificant in the case of the simple model because of the very low dimension numbers of the control variables and the observations.

Acknowledgments

We thank Si Shen for some valuable comments.

Author Contributions

Conceptualization: Bin Wang.

Formal analysis: Bin Wang, Juanjuan Liu.

Funding acquisition: Bin Wang, Juanjuan Liu.

Methodology: Bin Wang, Juanjuan Liu.

Software: Li Liu, Shiming Xu, Wenyu Huang.

Writing – original draft: Bin Wang, Juanjuan Liu.

Writing – review & editing: Bin Wang, Juanjuan Liu.

References

1. Evensen G, Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte-Carlo Methods to Forecast Error Statistics. *J. Geophys. Res.*, 1994; 99(C5), 10143–10162, https://doi.org/10.1029/94JC00572
2. Houtekamer PL and Mitchell HL, Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.* 1998; 126, 796–811. http://dx.doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2
3. Hamill TM, Whitaker JS and Snyder C, Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, 2001; 129, 2776–2790. http://dx.doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2
4. Kalnay E, Li H, Miyoshi TS, Yang C and BallabreraPoy J, 4-D-Var or ensemble Kalman filter? *Tellus A* 2007; 59: 758–773. https://doi.org/10.1111/j.1600-0870.200700261.x
5. Evensen G, *Data assimilation*, *The Ensemble Kalman Filter*. 2nd ed., Springer; 2009.
6. Anderson JL, Anderson SL, A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Mon. Wea. Rev.*1999; 127, 2741–2758. http://dx.doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2
7. Pham DT, Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems. *Mon. Wea. Rev.* 2001; 129, 1194–1207. http://dx.doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2
8. Wang XG, Bishop CH. A Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *J. Atmos. Sci.*, 2003; 60, 1140–1158. http://dx.doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2
9. Anderson JL, An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, 2007; 59: 210–224. https://doi.org/10.1111/j.1600-0870.2006.00216.x
10. Zhang F, Zhang M and Hansen JA, Coupling ensemble Kalman filter with four- dimensional variational data assimilation, Adv. Atmos. Sci. 2009; 26, 1–8. https://doi.org/10.1007/s00376-009-0001-8
11. Keppenne CL, Data assimilation into a primitive equation model with a parallel ensemble Kalman filter. *Mon. Wea. Rev.* 2000; 128, 1971–1981. http://dx.doi.org/10.1175/1520-0493(2000)128<1971:DAIAPE>2.0.CO;2
12. Anderson JL, An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.* 2001; 129, 2884–2903. http://dx.doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2
13. Houtekamer PL and Mitchell HL, A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 2001; 129, 123–137. http://dx.doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2
14. Whitaker JS and Hamill TM, Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.* 2002; 130, 1913–1924. http://dx.doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2
15. Ott E, Hunt BR, Szunyogh I, Zimin AV, Kostelich EJ, Corazza M., et al. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, 2004; 56: 415–428. https://doi.org/10.1111/j.1600-0870.2004.00076.x
16. Buehner M, Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Q.J.R. Meteorol. Soc.*, 2005; 131: 1013–1043. https://doi.org/10.1256/qj.04.15
17. Bishop CH and Hodyss D, Flow adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Q.J.R. Meteorol. Soc.*, 2007; 133: 2029–2044. https://doi.org/10.1002/qj.169
18. Constantinescu EM, Sandu A, Chai TF, and Carmichael GR. Ensemble-based chemical data assimilation. II: Covariance localization. *Q.J.R. Meteorol. Soc.*, 2007; 133: 1245–1256. https://doi.org/10.1002/qj.77

19. Fertig EJ, Hunt BR, Ott E, and Szunyogh I. Assimilating non-local observations with a local ensemble Kalman filter. *Tellus A* 2007; 59, 719–730. https://doi.org/10.1111/j.1600-0870.2007.00260.x

20. Oke PR, Sakov P. and Corney SP, Impacts of localization in the EnKF and EnOI: experiments with a small model. Ocean Dyn. 2007; 57, 32–45. https://doi.org/10.1007/s10236-006-0088-8

21. Liu CS, Xiao N, and Wang B. An Ensemble-Based Four-Dimensional Variational Data Assimilation Scheme. Part II: Observing System Simulation Experiments with Advanced Research WRF (ARW). *Mon. Wea. Rev*. 2009; 137, 1687–1704. http://dx.doi.org/10.1175/2008MWR2699.1

22. Wang B, Liu JJ, Wang S, Cheng W, Liu J, Liu CS, et al. An economical approach to four-dimensional variational data assimilation. Adv. Atmos. Sci. 2010; 27, 715–727. https://doi.org/10.1007/s00376-009-9122-3

23. Liu J, Wang B and Xiao Q, An evaluation study of the DRP-4-DVar approach with the Lorenz-96 model. *Tellus A*, 2011; 63: 256–262. https://doi.org/10.1111/j.1600-0870.2010.00487.x

24. Zhu J, Zheng F, and Li X, A new localization implementation scheme for ensemble data assimilation of non-local observations. *Tellus A*, 2011; 63: 244–255. https://doi.org/10.1111/j.1600-0870.2010.00486.x

25. Anderson JL, Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation. *Mon. Wea. Rev*. 2012; 140, 2359–2371. http://dx.doi.org/10.1175/MWR-D-11-00013.1

26. Sang H and Huang JZ, A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2012; 74: 111–132. https://doi.org/10.1111/j.1467-9868.2011.01007.x

27. Anderson JL, Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A* 2009; 61: 72–83. https://doi.org/10.1111/j.1600-0870.2008.00361.x

28. Gaspari G and Cohn SE, Construction of correlation functions in two and three dimensions. *Q.J.R. Meteorol. Soc*.1999; 125: 723–757. https://doi.org/10.1002/qj.49712555417

29. Miyoshi T and Sato Y, Assimilating Satellite Radiances with a Local Ensemble Transform Kalman Filter (LETKF) Applied to the JMA Global Model (GSM). *SOLA*, 2007; 3, 37–40. https://doi.org/10.2151/sola.2007-010

30. Campbell WF, Bishop CH, and Hodyss D, Vertical covariance localization for satellite radiances in ensemble Kalman filters. *Mon. Wea. Rev*. 2010; 138, 282–290. http://dx.doi.org/10.1175/2009MWR3017.1

31. Buehner M, Houtekamer PL, Charette C, Mitchell HL, and He B, Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. Mon. Wea. Rev. 2010a; 138, 1550–1566.

32. Buehner M, Houtekamer PL, Charette C, Mitchell HL, and He B, Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. Mon. Wea. Rev. 2010b; 138, 1567–1586.

33. Bishop CH, Hodyss D, Steinle P, Sims H, Clayton AM, Lorenc AC, et al. Efficient Ensemble Covariance Localization in Variational Data Assimilation. Mon. Wea. Rev., 2011; 139, 573–580.

34. Kuhl DD, Rosmond TE, Bishop CH, McLay J, and Baker NL, Comparison of Hybrid Ensemble/4DVar and 4DVar within the NAVDAS-AR Data Assimilation Framework. Mon. Wea. Rev., 2013; 141, 2740–2758.

35. Sakov P. and Oke PR. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. Tellus A 2008; 60: 361–371. https://doi.org/10.1111/j.1600-0870.2007.00299.x

36. Lorenz E, *Predictability: a problem partly solved*. In: Proc. Seminar on Predictability. Volume 1, reading, ECMWF, United Kingdom, 1996: pp.1–19.

37. Lawson WG, Hansen JA, Implications of Stochastic and Deterministic Filters as Ensemble-Based Data Assimilation Methods in Varying Regimes of Error Growth. *Mon. Wea. Rev*., 2004; 132, 1966–1981.

38. Lorenz E, and Emanuel K, Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci*. 1998; 55, 399–414. http://dx.doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2

39. Wang B and Ji ZZ, Construction and numerical tests of the multi-conservation difference scheme. *Chinese Science Bulletin*, 2003; 48 (10), 1016–1020.

40. Wang B and Tan XW, Conditional Nonlinear Optimal Perturbations: Adjoint-Free Calculation Method and Preliminary Test. *Mon. Wea. Rev*., 2010; 138, 1043–1049. http://dx.doi.org/10.1175/2009MWR3022.1

41. Bergemann K and Reich S, A localization technique for ensemble Kalman filters. *Q. J. R. Meteorol. Soc*. 2010; 136: 701–707. https://doi.org/10.1002/qj.591