

## Research Article

# The Impact of Length-Scale Variation When Diagnosing the Standard Deviations of Background Error in a 4D-Var System and Filtering Method Investigation

Xiang Xing <sup>1,2</sup>, Bainian Liu,<sup>1</sup> Weimin Zhang,<sup>1</sup> Xiaoqun Cao <sup>1</sup>, and Hongze Leng<sup>1</sup>

<sup>1</sup>College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China

<sup>2</sup>College of Computer, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Xiang Xing; [xingxiang14@nudt.edu.cn](mailto:xingxiang14@nudt.edu.cn)

Received 27 June 2020; Accepted 16 September 2020; Published 5 October 2020

Academic Editor: Helena A. Flocas

Copyright © 2020 Xiang Xing et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The four-dimensional variational data assimilation (4D-Var) method has been widely employed as an operational scheme in mainstream numerical weather prediction (NWP) centers. In addition to the ensemble data assimilation method, the randomization technique is still used to diagnose the standard deviations of background error in variational data assimilation (VAR) systems; however, such randomization techniques induce sampling noise, which may contaminate the quality of the standard deviations. First, this paper studies the properties of the sampling noise induced by the randomization technique. The results show that the sampling noise is on a small scale displaying high-frequency oscillations around the estimate compared with the estimate and this difference motivates the use of filtering techniques to eliminate the sampling noise effects. The characteristics of the standard deviation field of the control variables are also investigated, and the standard deviation fields of different model parameters have different scales and vary with the vertical model levels. To eliminate such sampling noise, the spectral filtering method used widely in the operational system and a modified spatial averaging approach are investigated. Although both methods have splendid performance in eliminating sampling noise, the spatial averaging approach is more efficient and easier to implement in operational systems. In addition, the optimal filtered results from the spatial averaging approach are dependent on model parameters and vertical levels, which is consistent with the variation in the standard deviation field. Finally, the spatial averaging approach is tested on the operational system at the global scale based on the YH4DVAR and the global NWP system, and the results indicate that the spatial averaging approach has positive effects on both analysis and forecast quality.

## 1. Introduction

A data assimilation system optimizes the combination of observations and a short-term forecast of the state variables to provide the best estimate of the initial state of the atmosphere [1]. Reports from the famous numerical weather prediction (NWP) Centre, European Centre for Medium-Range Weather Forecasts (ECMWF), show that the assimilated observations contribute 15% of the influence in any one analysis, and the complementary 85% is from the background which contains information from earlier assimilated observations [2]. The information from observations and a short-term forecast are combined and weighted by their corresponding error covariance matrix, the

observation error covariance matrix (**R** matrix), and the background error covariance matrix (**B** matrix). The four-dimensional variational data assimilation (4D-Var) method is the mainstream scheme used by many leading NWP centers to estimate the initial state of the atmosphere for the NWP model [3]. 4D-Var can assimilate observations of nontraditional types, in particular from satellites or radar, and these observations are generally considered to be a key factor in the continuous improvement of the quality of NWP [4–6]. In the 4D-Var system, the **B** matrix has a profound impact on the analysis because it determines the weight of the prior state, propagates the information from observations, and imposes a balanced relationship between the model control variables [7, 8].

The specification of the  $\mathbf{B}$  matrix is a key component of the 4D-Var system, but it remains a great challenge. The dimensions of the model state vector are of the order of  $10^7$  in a current NWP system; thus, it is far too large for the  $\mathbf{B}$  matrix to be explicitly represented. The lack of knowledge about the statistical properties of background error is also a limitation. To overcome these difficulties, significant efforts have been made to approximate the  $\mathbf{B}$  matrix. In the operational variational data assimilation (VAR) system, generally, the  $\mathbf{B}$  matrix has to be modeled by using control variable transforms (CVT) and represented as a product of operator form,  $\mathbf{B} = \Sigma \mathbf{C} \Sigma$ , where  $\mathbf{C}$  is the background error correlation operator and  $\Sigma$  is the diagonal matrix of background error standard deviations [9–14]. Considering the constraints between multiple variables, the standard deviations need to be diagnosed in the VAR system. Fisher and Courtier [15] suggested that the standard deviations of background error can be estimated by a randomization technique if the  $\mathbf{B}$  matrix is in the form  $\mathbf{B} = \mathbf{U} \mathbf{U}^T$ . The randomization approximation of the standard deviations is widely implemented in the current operational centers [16].

The YH4DVAR system is an operational deterministic 4D-Var system that uses the global spectral model as a constraint to impose a dynamic balance on the assimilation [17]. Considering that the sampling noise introduced by the randomization technique has a negative effect on the estimate, a spectral filtering method truncated at a fixed

wavenumber in an operational setting solves the issue of sampling noise in the YH4DVAR system [18]. The spectral filtering method enables the removal of most of the sampling noise while preserving the standard deviations of interest. However, the local variation characteristics of the standard deviations may be destroyed through fixed wavenumber truncation. In this paper, we will introduce the modified spatial averaging approach to solve the issue of sampling noise.

The article is organized as follows. First, the current operational version of the background error covariance matrix model in the YH4DVAR system is introduced in Section 2, together with the spatial structure of sampling noise and the characteristics of the local variation in the standard deviations for further investigations. In Section 3, filtering approaches are implemented on the operational system and the filtered results are obtained. The results of applying the filtering approaches to the operational system are presented in Section 4. Finally, conclusions follow in Section 5.

## 2. Motivations

In a 4D-Var system, the cost function measures the distance between the model trajectory and the observations over an assimilation window. The cost function [19] can be written in terms of increment  $\delta \mathbf{x}$  by

$$J_{\text{inc}}(\delta \mathbf{x}) = \frac{1}{2} (\delta \mathbf{x})^T \mathbf{B}_0^{-1} (\delta \mathbf{x}) + \frac{1}{2} \sum_{t=0}^T (\mathbf{d}_t - \mathbf{H}_t(\mathbf{M}_{0,t} \delta \mathbf{x}))^T \mathbf{R}_t^{-1} (\mathbf{d}_t - \mathbf{H}_t(\mathbf{M}_{0,t} \delta \mathbf{x})), \quad (1)$$

where subscript  $t$  is the time index,  $\delta \mathbf{x}$  is the increment, and  $\mathbf{M}_{0,t} \delta \mathbf{x}$  is the increment evolved by the tangent linear model from the initial time to time index  $t$ .  $\mathbf{B}$  and  $\mathbf{R}_t$  are the covariance matrices of background errors and observation errors at time index  $t$ , respectively.  $\mathbf{d}_t$  is the innovation vector at each time step,  $\mathbf{d}_t = \mathbf{y}_t^o - \mathcal{H}_t(\mathbf{x}_t^b)$ , where  $\mathbf{y}_t^o$  is the observation vector,  $\mathbf{x}_t^b$  is the background at time index  $t$ , and  $\mathcal{H}_t$  is the observation operator with a linear approximation  $\mathbf{H}_t$ .

To simplify the background term in  $J_{\text{inc}}$ , most leading NWP centers are modeling the  $\mathbf{B}$  matrix in the CVT framework. A general CVT form of a covariance matrix can be written as an eigenvector decomposition [20]. In the VAR system, the  $\mathbf{B}$  matrix can be modeled as  $\mathbf{B} = \mathbf{U} \mathbf{U}^T$ , which is not computed explicitly in the assimilation but implied by the operator  $\mathbf{U}$ . The construction of operator  $\mathbf{U}$  is a challenge in background error covariance modeling, and it is currently decomposed into the parameter transform  $\mathbf{U}_p$ , horizontal transform  $\mathbf{U}_h$ , and vertical transform  $\mathbf{U}_v$  in operational settings [21]:

$$\delta \mathbf{x} = \mathbf{U}_p \delta \mathbf{x}_u, \quad (2)$$

where  $\mathbf{U}_p$  introduces balance constraints to the assimilations and the parameters of  $\mathbf{x}_u$  are thought to be nearly uncorrelated

with each other [22–24]. Therefore, the background error covariance matrix can be modeled by  $\mathbf{B} = \mathbf{U}_p \mathbf{B}_u \mathbf{U}_p^T$ , where the matrix  $\mathbf{B}_u$  is the covariance of  $\mathbf{x}_u$ .

*2.1. Randomization Technique and Sampling Noise.* In the YH4DVAR operational system, the control variables contain the vorticity  $\zeta$ , the divergence  $\eta$ , the temperature and surface pressure  $(T, p_s)$ , and the specific humidity  $q$ . The  $\mathbf{U}_p$  matrix transforms  $[\zeta, \eta_u, (T, p_s)_u, q_u]$  into  $[\zeta, \eta, (T, p_s), q]$ , and the parameters of  $\mathbf{x}_u$  correspond to the unbalanced components of the control variables; thus,  $\mathbf{B}_u$  is assumed to be a block-diagonal matrix with no correlation between the parameters. Therefore, the background error covariance matrix can be written explicitly by the autocovariance matrix  $\mathbf{B}_u$  and parameter transform  $\mathbf{U}_p$ ,  $\mathbf{B} = \mathbf{U}_p \mathbf{B}_u \mathbf{U}_p^T$ , with the background autocovariance matrix  $C(\cdot)$  for each variable. In the VAR system, it is crucial to input the standard deviations of the background error of control variables into the autocovariance matrix  $C(\cdot)$  to estimate the  $\mathbf{B}$  matrix, and the standard deviations of other variables are implicitly given by the parameter transform.

Fisher and Courtier suggested that the randomization technique can be used to diagnose the standard deviations of background error in the assimilation system, thereby

alleviating the impact of the imposition of balance conditions on the correlations of background error [15]. As the  $\mathbf{B}$  matrix is of the form  $\mathbf{B} = \mathbf{U}\mathbf{U}^T$ , the standard deviations of background error can be estimated by the sample standard deviations of model vectors, that is, applying  $\mathbf{U}$  to a set of  $N$  ( $N=10$ ) random vectors  $\phi_1, \phi_2, \dots, \phi_n$  drawn from a Gaussian distribution with  $N(0, I)$ . Figure 1 shows the vorticity standard deviation field at approximately 1000 hPa obtained by the randomization technique. Figures 1(a)–1(f) correspond to the raw estimates obtained with 10, 50, 100, 200, 1000, and 10000 samples, respectively, which are abbreviated as  $\zeta(10)$ ,  $\zeta(50)$ ,  $\zeta(100)$ ,  $\zeta(200)$ ,  $\zeta(1000)$ ,  $\zeta(10000)$ . The sampling noise tends to be characterized by a relatively small scale in  $\zeta(10)$  such that the estimate strongly deteriorated, and the sampling noise is significantly reduced when the sample size increases to 50 that the main large-scale features of the vorticity standard deviation field are well captured. The values of the vorticity estimate are smaller in the tropics and data-dense areas and higher in the high latitude areas, especially in the southern hemisphere in  $\zeta(50)$ . As the number of samples increases, the sampling noise is dramatically removed, and the scale of the standard deviation field is visible. When the number of samples reaches 1000, the basic shape of the estimate ceases to meaningfully change; for example,  $\zeta(1000)$  is similar to  $\zeta(10000)$ .

The randomization technique is widely used to diagnose the standard deviations of background error in VAR systems in many operational centers, and the sampling noise is inevitably introduced into the estimate when undersampling. Although increasing the sample number can alleviate this issue, it is usually unrealistic to employ a sample size over 100 for estimation in the operational system. It needs to be a compromise between the accuracy of the estimates and operational time constraints that considers maintaining affordable computational costs in the operational system. Therefore, it is important to remove the sampling noise efficiently in the case of undersampling.

To separate the sampling noise induced by the randomization technique from the estimate, it is necessary to investigate the difference between the sampling noise and the standard deviation field. The spatial structure of the sampling noise in the assimilation system will be explored in the operational version of the  $\mathbf{B}$  matrix. The true state of background error standard deviations for vorticity is never exactly known, but it is known from Figure 1 that when the number of samples reaches 1000, the estimate of the vorticity field is sufficiently accurate. Thus, the reference for the vorticity standard deviation field is generated from 1000 samples. The sampling noise included in the estimates is obtained by subtracting the reference  $\zeta(1000)$  from the estimated values of 50, 100, and 200 samples. The sampling noise is defined by

$$\text{SN} = \zeta(i) - \zeta(1000), \quad (3)$$

The energy spectrum of the signal is an effective means to examine the spatial structure of the signal. The wavenumber in the spectral space corresponds to the scale of the spatial

structure in the grid-point space that a large-scale spatial structure means that the energy of the small wavenumber in the spectral space is much larger than that of the large wavenumber, whereas it is a small scale. Figure 2 shows the sampling noise and the corresponding energy spectrum of a specific transect on the spatial fields from the 50 samples, 100 samples, and 200 samples. There are two striking features worth noting in Figure 2. First, the sampling noise is clearly presented in the vorticity standard deviation field and displays high-frequency oscillations around the globe on a small scale. From the corresponding energy spectrum in the right column, the energy spectrum of the noise over the entire band is more consistent and is close to the white noise. Another striking feature is that the structure distribution of sampling noise from different samples is similar, but as the number of samples increases, the absolute extreme value of the noise decreases, which is consistent with the fact that more samples generate a more accurate estimate. The information in Figures 1 and 2 indicates that the spatial structure of the sampling noise is different from that of the estimates, and this difference motivates the reduction in the sampling noise by filtering techniques.

*2.2. Characteristics of the Standard Deviation Field.* The characteristics of the standard deviations of background error obtained by the randomization technique in the assimilation system will be explored in the operational version of the  $\mathbf{B}$  matrix.

The characteristics of the standard deviation estimate vary with the model parameters, as illustrated in Figure 3. An overview of Figure 3 shows that low values of vorticity, divergence, temperature, and  $U$ -wind estimates are found in the tropics, and higher values are found in the high latitude areas, especially in the southern hemisphere. It is also shown that the standard deviation fields of the four model parameters have different distributions: low values of temperature and divergence are more concentrated at low latitudes, and higher values are found in higher latitude areas; however, there are relatively large vorticity fields along the intertropical convergence zone, and  $U$ -wind has a larger value in the data-spare and Antarctic regions and a smaller value in the densely observed land area.

The temporal evolution of the standard deviation field is examined in Figure 4 at 1200 UTC and 2400 UTC. It appears that the evolution of the standard deviations over 12 hours is not quite significant, and the distribution of the temperature standard deviation field at 500 hPa (1000 hPa), stretching along the west-east axis, is quite similar over the period. In the VAR system, the covariances matrix is estimated by the climatic information; thus, the characteristics of the standard deviation field obtained by the randomization technique have a certain similarity. In addition, the distribution of the standard deviation field is connected with the dynamics of the synoptic situation. There is another striking feature in Figure 4 in the differences in the distribution of the standard deviation field at the vertical levels of the model. The values near the surface are larger than those at 500 hPa, which can be explained by the complicated physical process

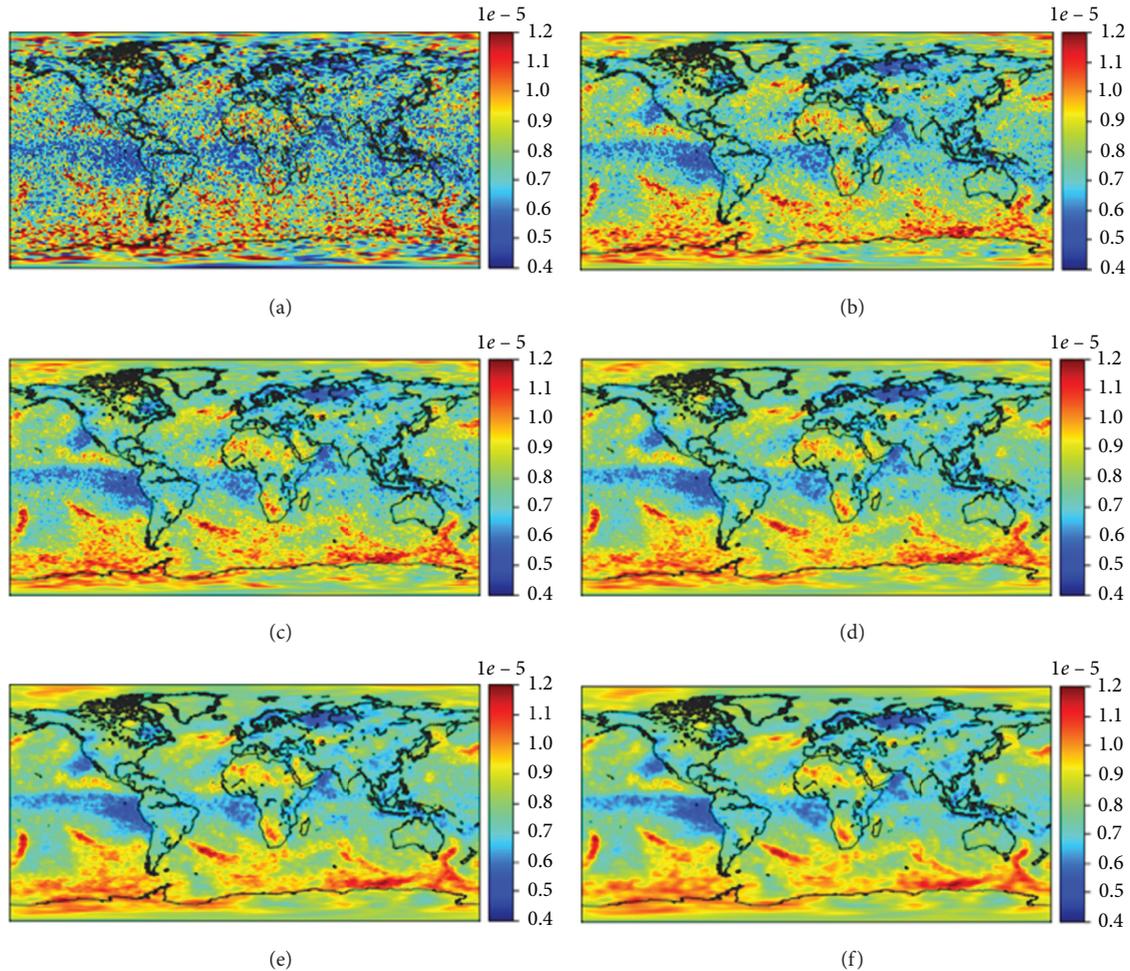


FIGURE 1: The vorticity standard deviation field at 1000 hPa estimated from (a) 10 samples, (b) 50 samples, (c) 100 samples, (d) 200 samples, (e) 1000 samples, and (f) 10000 samples.

near the surface directly affecting the relative error statistics. The corresponding energy spectra of the standard deviation at 500 hPa and 1000 hPa are plotted in Figure 5. Both of the standard deviations at different model levels have relatively larger energy at large scales, and the energy spectrum of the standard deviation at 1000 hPa has a larger amplitude at medium scales than that at 500 hPa.

The length-scale of the estimate of the standard deviations varies with the model parameters and the vertical levels, and these variations have a profound impact on the procedure for eliminating sampling noise. Therefore, the method of removing the sampling noise should not only preserve the standard deviations of interest but also consider the variation in the standard deviation field with model parameters and levels.

### 3. Filtering Approaches

The standard deviations of background error estimated by the randomization technique are contaminated by sampling noise, and this noise has a negative effect on the estimate. The previous paragraph shows that the sampling noise differs in the spatial structure based on the standard

deviations and that the standard deviations are model parameter-dependent and vertical level-dependent. In this section, the spectral filtering method used in the operational system and the concept of a spatial averaging approach will be introduced, with a discussion of the implementation of these approaches in the YH4DVAR system. The filtered results of spatial averaging on the estimate will also be discussed in detail.

**3.1. Filtering Approaches and Implementation.** In the YH4DVAR operational system, a spectral filtering method, truncated at a fixed wavenumber  $N_{tr}$ , is employed to remove the sampling noise by multiplying a coefficient  $\rho(n)$  with each wavenumber. The coefficient  $\rho(n)$  is defined by

$$\rho(n) = \left[ \cos\left(\frac{\pi \min(n_s, N_{tr})}{2 N_{tr}}\right) \right]^2, \quad (4)$$

where  $n_s$  is the wavenumber in the spectral space. The spectral filtering method achieves the elimination of sampling noise by setting a reasonable truncated wavenumber  $N_{tr}$ , based upon empirical settings, and the efficiency of the

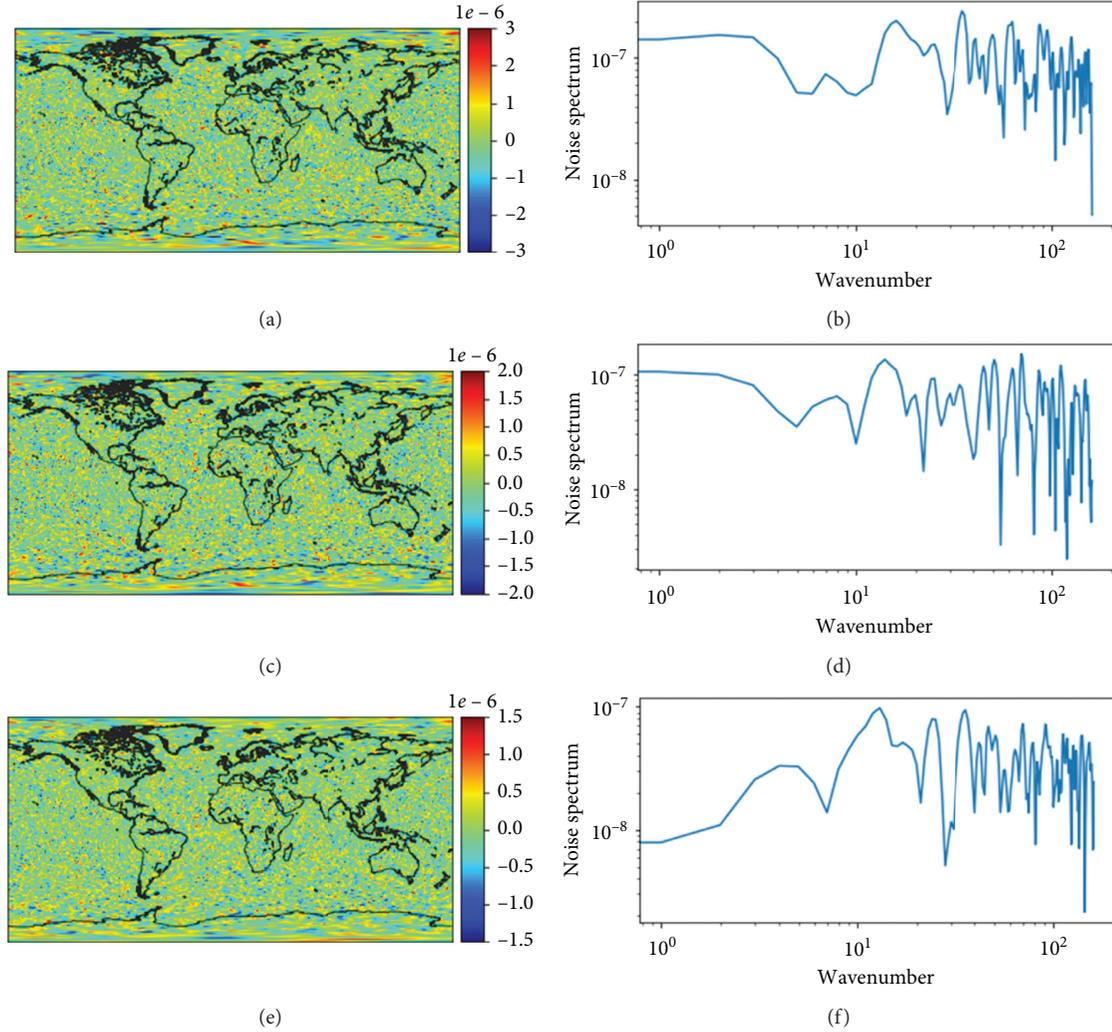


FIGURE 2: The noise of the vorticity standard deviation field at 1000 hPa calculated from 50 samples and the corresponding energy spectrum (first row), 100 samples (second row), and 200 samples (third row).

spectral filtering method is more significant when the standard deviation field is isotropic and varies slowly.

For the filtered results in grid-point space, the spatial averaging [25] of the estimate is defined by

$$\sigma_f(i) = \frac{1}{N_x * N_y} \sum_{lx=-L/2}^{L/2} \sum_{ly=-L/2}^{L/2} \sigma_{(lx,ly)}, \quad (5)$$

where  $N_x = n_L + 1$ ,  $L$  is the length of spatial averaging, and  $n_L$  is the number of grid points in domain  $[-L/2, L/2]$ . The double sum in this equation indicates that spatial averaging has a potential effect of increasing the sampling realizations. This means that, in each model level, the total sampling realizations will be dramatically increased by  $N_t = N \times N_x \times N_y$ , where  $N$  is the sampling number. Thus, the averaged results have less sampling noise when the independent sampling realizations are increased.

It is important to investigate the implementation and computational cost of filtering approaches when considering their application to operational systems. The implementation of the spectral filtering and spatial averaging approach

on the operational system is discussed, and the flowchart is shown in Figure 6. The practical implementation of the spectral filtering method using fixed wavenumber truncation in the operational system can be summarized by the following steps:

- (1) For a set of  $N$  ( $N=50$ ) random vectors  $\phi_1, \phi_2, \dots, \phi_n$  drawn from a Gaussian distribution with  $N(0, I)$ , apply  $U$  to each vector to obtain a set of model vectors  $b_1, b_2, \dots, b_n$  in spectral space, in which  $b_i = U\phi_i$
- (2) Compute the standard deviations after transforming the model parameters into grid-point space
- (3) Transform the standard deviations into spectral space, then multiply the coefficient  $\rho(n)$
- (4) Transform the filtered standard deviations into grid-point space

Spatial averaging can be performed directly in the grid-point space to obtain the filtered results after the standard deviations are obtained in the second step above; therefore,

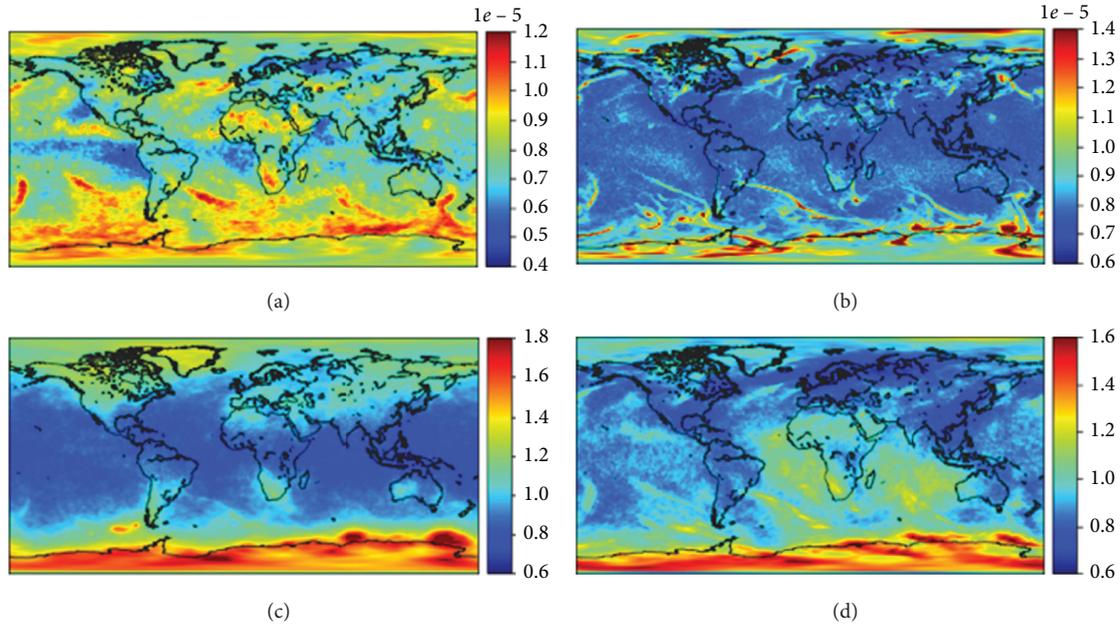


FIGURE 3: (a) Vorticity, (b) divergence, (c) temperature, and (d)  $U$ -wind standard deviation field at 1000 hPa.

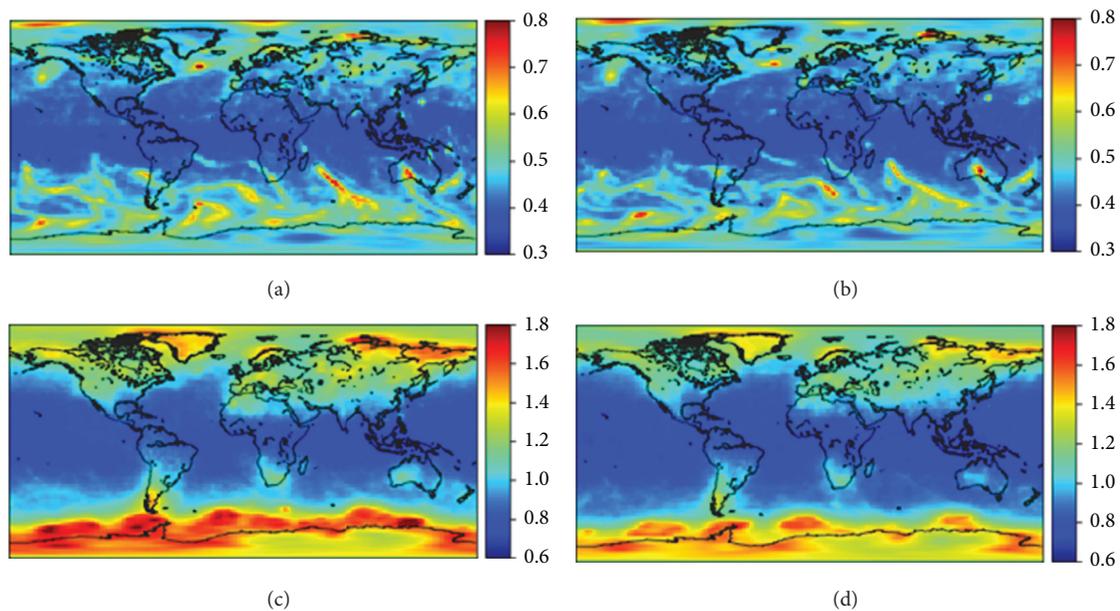


FIGURE 4: Temperature standard deviation field at 500 hPa (a) at 1200 UTC, (b) at 2400 UTC, and (c) and (d) as the former but at 1000 hPa.

the spatial averaging approach does not need to transform standard deviations into spectral space and has a higher degree of parallelism in the grid-point space. There is another striking feature that once the optimal spatial average length is determined, there is no need to relocate the grid points satisfying the distance because the sample realizations in the field are invariable. The calculation of the optimal spatial averaging length  $L$  can be performed offline and updated with the frequency of the background error covariance statistics.

**3.2. Filtering Effects.** Figure 7(a) is the reference for the vorticity standard deviation field generated from 1000 samples, and the raw estimate computed from 50 samples is represented in Figure 7(b). It is possible to capture the main large-scale features with only 50 samples, and the sampling noise tends to be characterized by a relatively small scale, such that the details of the estimate strongly deteriorated. Note that in the operational assimilation system, it is usually unrealistic to employ a sample size over 100 for estimating because of the computational cost. The filtered result of the

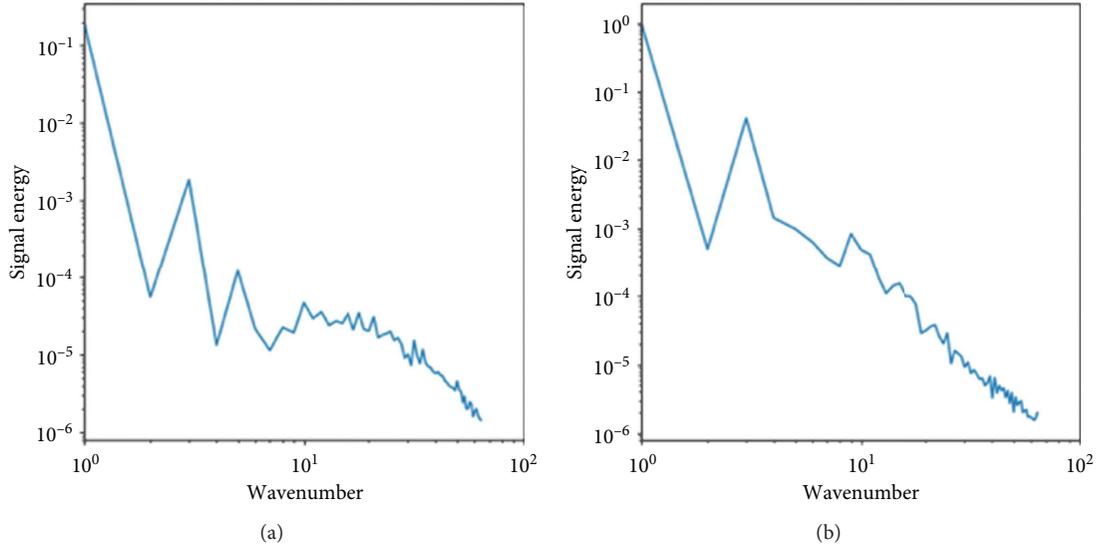


FIGURE 5: Energy spectra of the temperature standard deviation at (a) 500 hPa and (b) 1000 hPa.

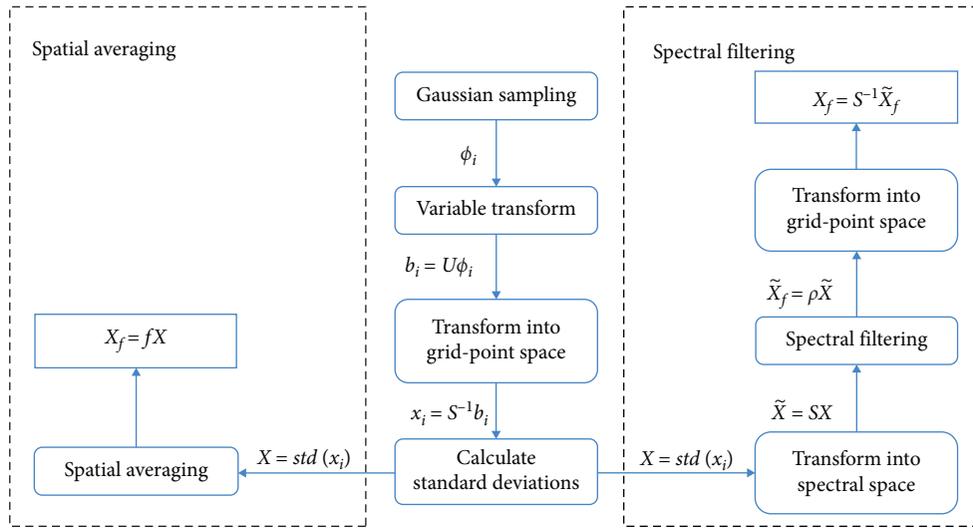


FIGURE 6: The flowchart of the standard deviation diagnostic procedure.

50 samples is illustrated in Figures 7(c) and 7(d) and are abbreviated as  $\zeta(50)_{\text{Ntr}}$  and  $\zeta(50)_{\text{sp}}$ , respectively.

Compared to Figure 7(b),  $\zeta(50)_{\text{Ntr}}$  and  $\zeta(50)_{\text{sp}}$  show that the sampling noise is weakened in the averaged results, and the scale of the standard deviation is visible. The signal-to-noise ratios of  $\zeta(50)_{\text{Ntr}}$  and  $\zeta(50)_{\text{sp}}$  are 23.74 and 25.41, respectively, both of which are larger than that of  $\zeta(50)$ , and a larger signal-to-noise ratio of  $\zeta(50)_{\text{sp}}$  strongly supports the fact that the spatial averaging is able to remove most of the detrimental noise.  $\zeta(50)_{\text{Ntr}}$  still has obvious noise, which is caused by the fixed truncation of the wavenumber. Figure 8 shows the coefficient of the fixed truncation of the wavenumber of the spectral filtering method. The  $N_{\text{tr}} = 1000$  used here is a conservative spectral filtering function. It sets most of the wavenumber with a

coefficient equal to 1, which mainly cuts off some low-frequency noise and does not consider the local variation of the standard deviation field. The physical structure captured by  $\zeta(50)_{\text{sp}}$  is very detailed and is even as accurate as the estimate in  $\zeta(1000)$ . The averaged results have less sampling noise as the independent sampling realizations are increased,  $N_t = N \times N_x \times N_y$ , which is coherent with the fact that the total sample size  $N_t$  of the filtered 50 samples is close to the sample size of the raw 1000 samples. This attractive result shows the potential efficiency of the spatial averaging in that the averaging estimates can reach the accuracy of large sample estimates by significantly eliminating the noise.

Figure 9 shows the relative error of the filtering result of vorticity  $\zeta(50)_{\text{sp}}$ , temperature  $T(50)_{\text{sp}}$ , and  $U$ -wind  $U(50)_{\text{sp}}$  at 1000 hpa:

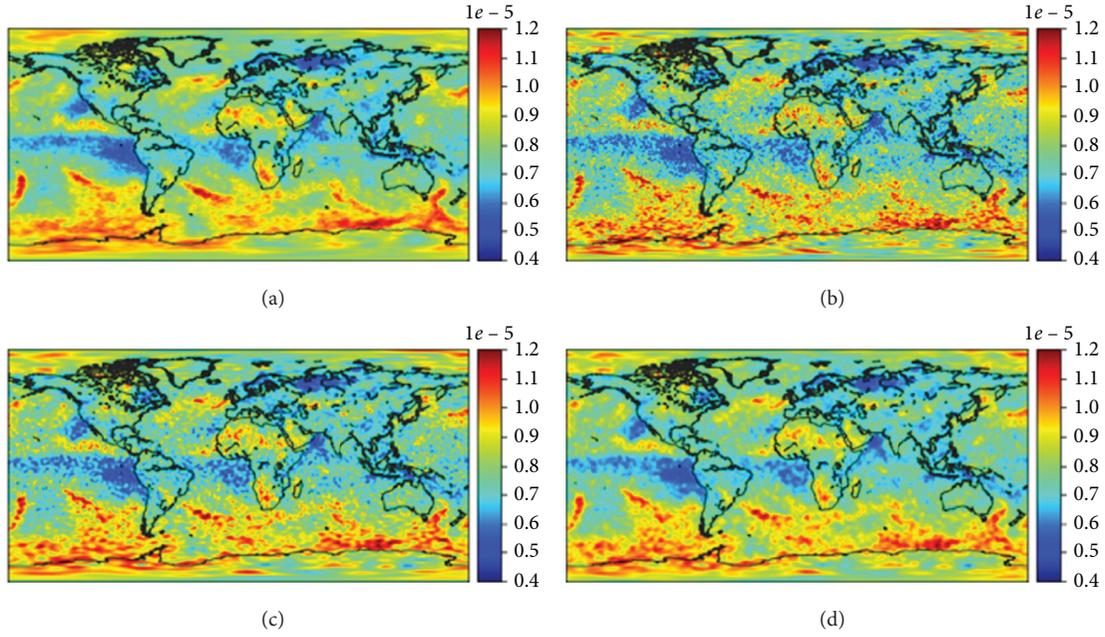


FIGURE 7: The vorticity standard deviation field at 1000 hPa estimated from (a) 1000 samples, (b) 50 samples, (c) filtered 50 samples with  $N_{tr} = 1000$ , and (d) spatially averaged 50 samples with the averaging length  $L = 500$  km.

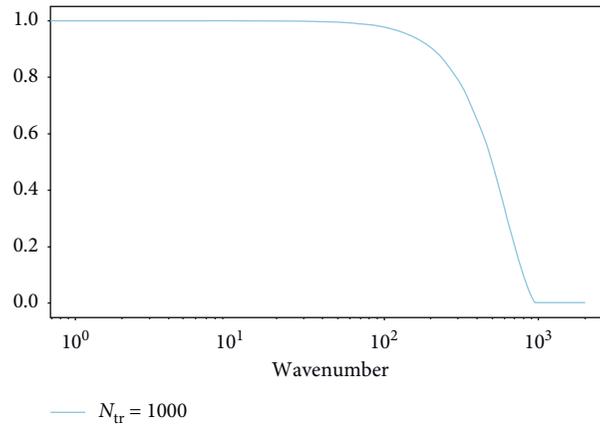


FIGURE 8: The coefficient of the spectral filtering method, truncated at  $N_{tr} = 1000$ .

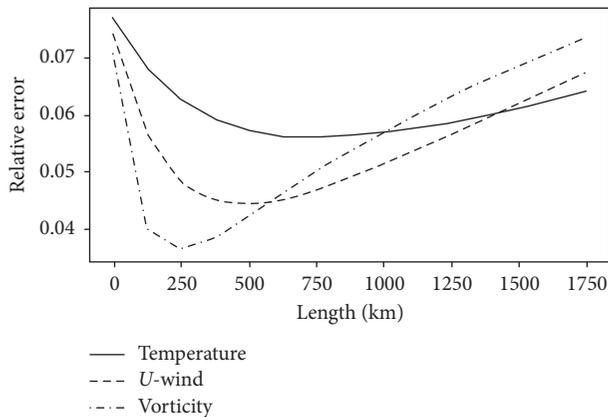


FIGURE 9: Relative error in the estimates of vorticity, temperature, and  $U$ -wind with different averaging lengths.

$$\text{relative error} = \left| \frac{\text{value}_{sp} - \text{value}_{ref}}{\text{value}_{ref}} \right|, \quad (6)$$

where  $\text{value}_{sp}$  is the standard deviation filtered by the spatial averaging approach and  $\text{value}_{ref}$  is the reference standard deviation field by 1000 realizations. It can be seen that, with the increase in the filtering length, the relative error of the filtering result can be expressed as a  $U$  shape. There is a significant effect on reducing the relative error as the averaging length increases, and there is an optimal length at which the error of the filtering result is minimized. When the averaging length becomes excessive, the relative error increases again with a greater length. This can be explained by spatial filtering increasing the accuracy of the filtered results by increasing the independent sample realizations; when the length becomes excessive, the relevant points are also

included in the realizations that result in a reduction in the filtering result. The optimal filtering length for vorticity is 250 km, for temperature is 750 km, and for  $U$ -wind is 500 km. The difference in the optimal length of these model parameters shows the difference in the standard deviation field, which is consistent with the characteristics of the standard deviation field mentioned in 2.2. It is absolutely necessary to consider the effects of the model parameters on the spatial averaging length in the procedure for removing sampling noise since the optimal filtering length is dependent on model parameters.

The efficiency of spatial averaging at the different vertical levels for different model parameters is shown in Figure 10. Figure 10(a) shows the optimal averaging length is different at the different model levels; for example, the optimal length of temperature around the surface is 750 km, which is larger than 500 km at 500 hPa. In addition, the trend in the optimal length varying with the vertical levels is different from that for model parameters. The optimal length of the  $U$ -wind increases obviously with the model levels, while the optimal length of the temperature decreases gradually from 1000 hPa to 500 hPa and then increases with the vertical model levels. In general, the optimal length of the variable increases with the vertical levels. This can be explained by the fact that the correlation length-scale of the variable field at higher levels is greater than that at the surface, and a greater length-scale corresponds to a greater optimal length for more independent sampling realizations. Figure 10(b) shows the relative error of  $U$ -wind raw estimates  $U(50)$  and optimal filtering results  $U(50)_{sp}$  at different vertical levels. The relative error of raw estimates  $U(50)$  remains around 0.08 while there is a significant reduction in the relative error with the optimal length. The essence of the spatial averaging to reduce sampling noise is to increase the independent sampling realizations by optimal length. A larger optimal length may result in the fact that the sampling realizations are not completely independent, and this is coherent with the relative error associated with optimal length being increasing above level 64.

#### 4. Effects on Analysis and Forecast Quality

The previous results show that the large-scale features of background error standard deviation can be captured by a randomization technique, but this technique may be contaminated by the sampling noise because of the finite sample sizes. Furthermore, the fixed wavenumber truncation in the spectral filtering method has an inferior impact on the local variation characteristics of the estimates, but the spatial averaging approach offers some advantages in addressing this issue. In this section, the effect of the spatial averaging approach on assimilation and forecast quality will be examined in the operational NWP system consisting of YH4DVAR and the global spectral model.

Applications for assimilation diagnostics and forecasting will be considered with the operational NWP system on a low-resolution grid with 91 levels (abbreviated as T399 L91).

There are two groups. In one, the number  $N$  of random vectors is taken as 50 in the first group, which is numbered CN, using the spectral filtering method truncated with a fixed wavenumber to eliminate the sampling noise. In the other experimental group, the number  $N$  of random vectors is taken as 10 with spatial averaging, numbered SP. The results conducted with the operational system on a high-resolution grid with 91 levels (abbreviated as T799 L91) are used as a reference. The operational system run started at 00 UTC on 26 August 2015, ended at 00 UTC on 26 September 2015, and was updated every 12 hours. The statistical properties of the  $B$  matrix are generated using the popular National Meteorological Centre (NMC) method, an approach that uses the differences between forecasts of different lengths to evaluate forecasts error [9], and the optimal averaging length of the model parameters is calculated at 00 UTC on 26 August 2015 on each vertical level.

Figure 11 displays the root mean square error (RMSE) of the geopotential height as a function of time for two schemes: CN and SP. The Northern and southern hemisphere extratropics correspond to the regions of  $30^{\circ}\text{N}$ – $90^{\circ}\text{N}$  and  $30^{\circ}\text{S}$ – $90^{\circ}\text{S}$ , respectively. The RMSE in the southern hemisphere is generally higher than that in the northern hemisphere and tropics, and RMSE in the tropics is slightly smaller than that in the northern hemisphere. The RMSE of the SP analysis in all three regions was less than that of the control, which leads to the same conclusion as that for 850 hPa in Figure 11(b). It can be observed that the RMSE in different regions is significantly reduced by the use of the spatial averaging approach in Table 1, and this tends to be consistent with the result that the spatial filtering has a positive impact on improving the accuracy of the estimate.

The difference between the SP and the CN analysis fields at a level near 500 hPa and 850 hPa is shown in Figure 12. A visible common feature is that the differences in low latitudes were significantly smaller than those in high latitudes. This can be interpreted to mean that the equatorial waves coupled to convection explain most of the standard deviation in the tropics, especially in the lower troposphere; thus, the filtering effect of the spatial averaging approach is similar to that of the spectral method. Another interesting point is that the larger values are mainly located in the vicinity of ridges and troughs, which is in agreement with previous studies.

Applications for prediction have been achieved by examining the forecast quality of filtered results on a global spectral model. Figure 13 shows the RMSE of the forecast for temperature (500 hPa) at different times compared with the references at the corresponding time of T799. The RMSE has a small value in the tropics and northern hemispheres, whereas much higher values are found in the southern hemispheres in the data-poor areas. It is also observed that the RMSE in the three regions increases with a long forecast time, and the growth of RMSE in the tropics is slower than that in the nontropical area. The filtered estimates have a positive impact on the forecast quality, and its forecast improvement is generally weaker than that in the analysis. This may be due to the filtering effect of the numerical model, and continued research is important to improve the skill of the forecast.

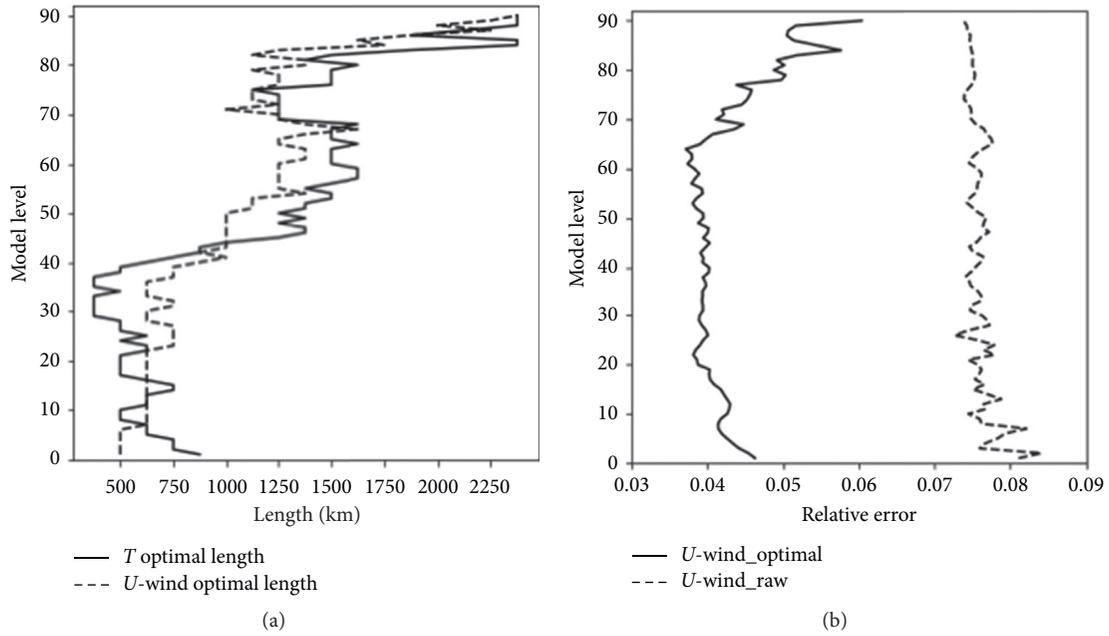


FIGURE 10: (a) The optimal length for temperature and *U*-wind by the vertical model level. (b) The relative error of *U*-wind raw estimates and optimal filtering results.

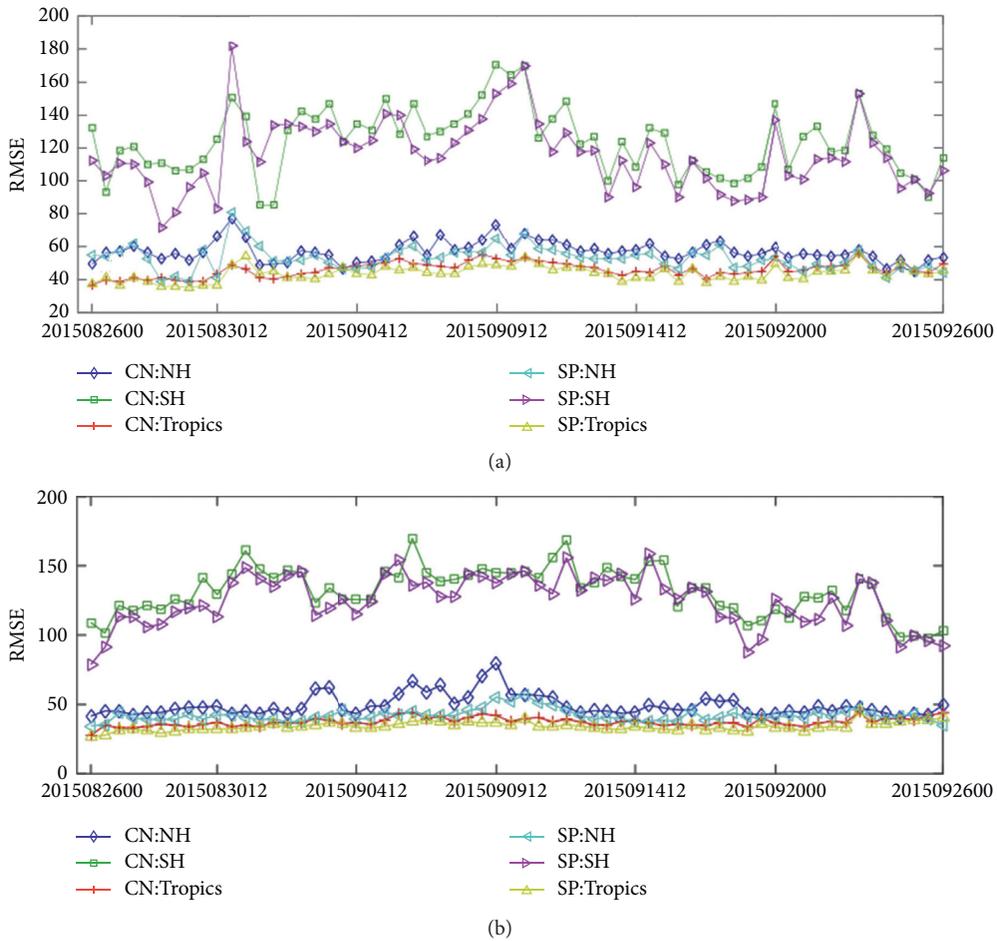


FIGURE 11: The RMSE of the 500 hPa geopotential height in different regions. (b) is the same but at 850 hPa.

TABLE 1: The monthly mean RMSE for 500 hPa and 850 hPa geopotential heights.

| Height (hPa) | Northern hemisphere |         | Southern hemisphere |          | Tropics |         |
|--------------|---------------------|---------|---------------------|----------|---------|---------|
|              | CN                  | SP      | CN                  | SP       | CN      | SP      |
| 500          | 57.0029             | 52.8616 | 124.1658            | 116.0676 | 46.0208 | 42.6178 |
| 850          | 49.2861             | 42.1323 | 131.9647            | 124.7308 | 37.4340 | 34.2293 |

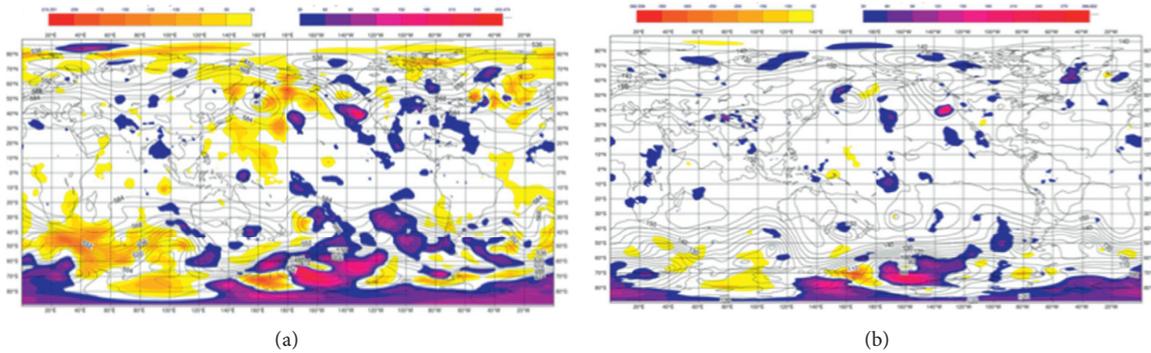


FIGURE 12: (a) The difference between SP and CN analysis at 500 hPa geopotential height. (b) is the same but at 850 hPa.

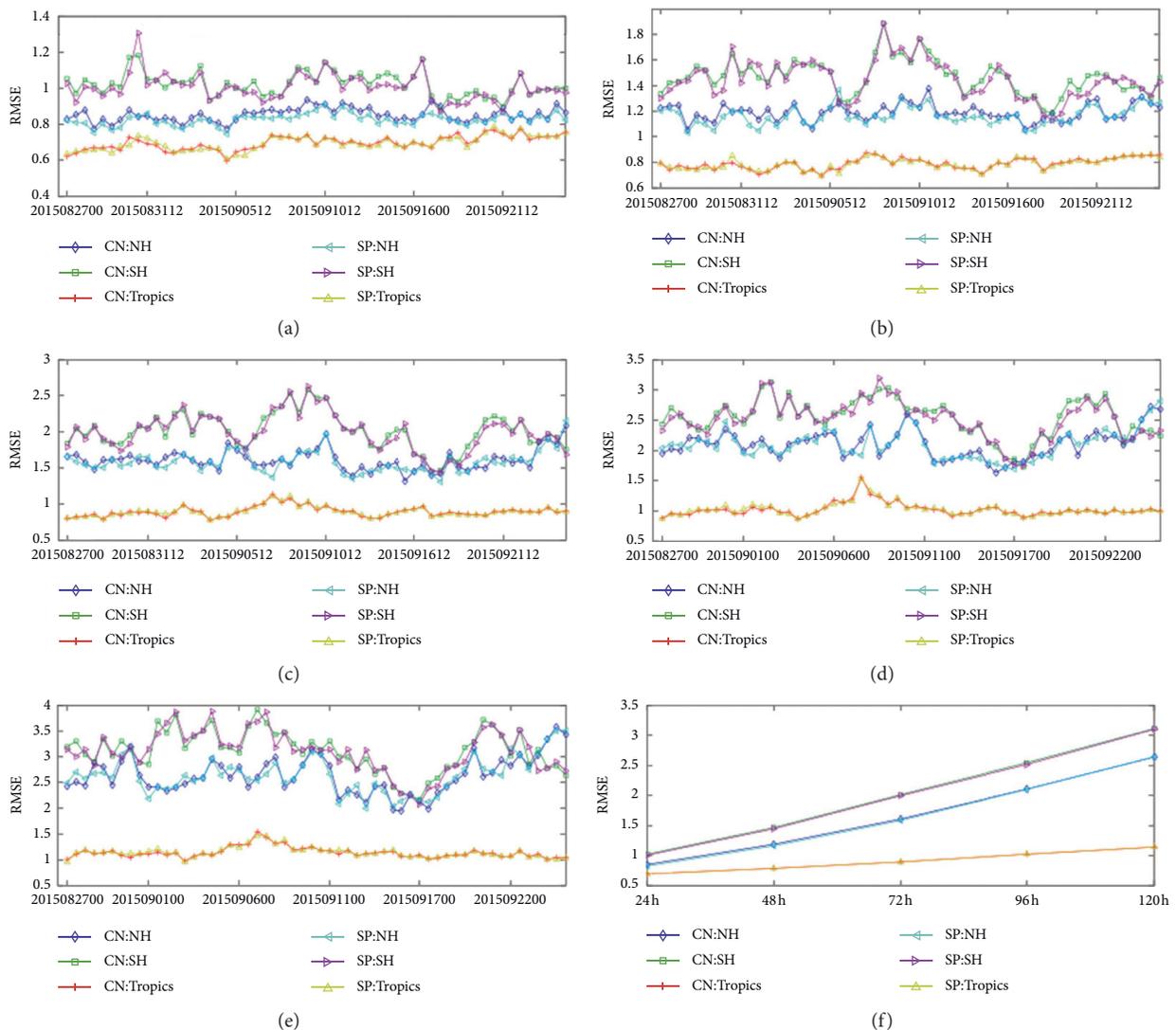


FIGURE 13: (a) RMSE of temperature versus 24-hour forecast time at 500 hPa. (b–e) RMSE of temperature versus 48 h, 72 h, 96 h, and 120 h forecast time at 500 hPa, respectively. (f) The mean RMSE.

## 5. Discussion and Conclusions

In a VAR system, the randomization technique is widely implemented in the current operational centers to diagnose the standard deviations of background error for modeling the **B** matrix. In this case, the standard deviations are estimated from a finite sample and will inevitably be affected by sampling noise because of the finite size of the samples. Therefore, some filtering techniques must be implemented to eliminate this sampling noise.

The spatial structure of the sampling noise introduced by the randomization technique is explored in the operational version of the **B** matrix. The estimated standard deviations from a finite sample are deteriorated by sampling noise, and the noise tends to be characterized by a relatively small scale and evenly distributed across the global map compared to the estimated reference field. The absolute value of noise dramatically decreases as the sample size increases, and the large-scale of standard deviations is more visible in the estimated field while the global distribution of sampling noise remains even. Such a difference in the structure between sampling noise and the estimates is the key to designing filtering techniques to filter out the small scale noise and preserve the signal of interest. There is a spectral filtering method truncated at a fixed wavenumber in an operational setting to resolve the issue of sampling noise in the YH4DVAR system. The application of this truncated spectral method is a kind of low-pass filter in spectral space that enables the small scale noise to be filtered out, preserves the large-scale information, and works well for eliminating sampling noise in the operational system.

Our investigation has also shown that the length-scale of the estimate of the standard deviations varies with the model parameters. The low values of the divergent standard deviations are more concentrated at low latitudes, and higher values are found in higher latitude areas, but there are relatively large vorticity fields along the intertropical convergence zone. In addition, the distribution of the standard deviation field is different between model vertical levels. The length-scale of the estimates from the randomization technique varies with the model parameters and vertical levels; thus, the fixed wavenumber truncation in the spectral filtering method has an inferior impact on the local variation characteristics of the standard deviations. The modified spatial averaging approach has some advantages in addressing this issue. In an operational context, the physical structure captured by the averaged results of 50 samples is fine, as the reference with the noise is dramatically eliminated, the spatial averaging approach has the potential to increase the independent sample realizations by averaging length, and the averaged results with the optimal length maximize the total sampling numbers. The optimal averaging length is connected with the model parameters and vertical level, which is consistent with the variation in the estimated standard deviations field; thus, the spatial averaging approach has better performance in eliminating the noise and preserving the characteristics of local variation in the estimates. In addition, the spatial

averaging approach is easy to implement in the operational system and has a higher degree of parallelism in the grid-point space. The RMSE of the analysis in different regions is significantly reduced by the use of the spatial averaging approach, and relatively small values can be observed over data-spare regions such as the southern hemisphere. In addition, the results also show that the spatial averaging approach has a positive effect on improving forecast skill scores.

In the next step, we need to make further improvements to the spatial averaging approach. The optimal averaging length is computed offline, and the averaging length could be suboptimal with the variations in flow-dependent information in the assimilation period. It would be interesting to consider updating the optimal averaging length adapted with the variations. In addition, some qualitative analysis conclusions given in this paper need to be examined in detail in future work. The homogenous and uniform filtering type of spatial averaging approach ignores the anisotropic characteristics of the signal; thus, some important characteristics may be smoothed or weakened in averaging, and it would be worth the effort to introduce an adaptive averaging algorithm in the future.

## Data Availability

The observations used in this paper are obtained from the NCEP and NOAA website (<ftp://ftp.prdd.ncep.noaa.gov/pub/data/nccf/com/gfs/prod>). The background data are not freely available because it is a short-term forecast by a global spectral model NWP system at the National University of Defense Technology that is subject to commercial confidentiality.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 41675097, 41375113, and 41605070) and the National Key R&D Program of China (2018YFC1506704).

## References

- [1] R. H. Reichle, "Data assimilation methods in the Earth sciences," *Advances in Water Resources*, vol. 31, no. 11, pp. 1411–1418, 2008.
- [2] C. Cardinali, S. Pezzulli, and E. Andersson, "Influence-matrix diagnostic of a data assimilation system," *Quarterly Journal of the Royal Meteorological Society*, vol. 130, no. 603, pp. 2767–2786, 2004.
- [3] F. Rabier, "Overview of global data assimilation developments in numerical weather-prediction centres," *Quarterly Journal of the Royal Meteorological Society*, vol. 131, no. 613, pp. 3215–3233, 2006.

- [4] X.-Y. Huang and D. M. Barker, "Four-dimensional variational data assimilation for WRF: formulation and preliminary results," *Monthly Weather Review*, vol. 137, no. 1, pp. 299–314, 2009.
- [5] J. R. Xiao, "Progress achieved on assimilation of satellite data in numerical weather prediction over the last 30 years," in *Proceedings of the ECMWF Seminar on Recent Development in the Use of Satellite Observations in NWP*, pp. 1–28, New York, NY, USA, 2008.
- [6] Q. Chongqin, Z. Lei, and S. Aimei, "An explicit four-dimensional variational data assimilation method," *Science China Earth Sciences*, vol. 50, no. 8, pp. 1232–1240, 2007.
- [7] R. Daley, *Atmospheric Data Analysis*, Cambridge University Press, Cambridge, UK, 1993.
- [8] F. Bouttier and P. Courtier, "Data assimilation concepts and methods march," 2002.
- [9] D. F. Parrish and J. C. Derber, "The national meteorological center's spectral statistical-interpolation analysis system," *Monthly Weather Review*, vol. 120, no. 8, pp. 1747–1763, 1992.
- [10] M. Fisher, "Wavelet Jb-A new way to model the statistics of background errors," *ECMWF Newsl.* vol. 200, no. 106, pp. 23–28, 2006.
- [11] M. Fisher, "Generalized frames on the sphere, with application to the background-error covariance modelling," in *Proceedings of the Seminar on Recent Developments in Numerical Methods for Atmospheric and Ocean Modelling*, pp. 87–101, New York, NY, USA, 2004.
- [12] R. N. Bannister, "A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances," *Quarterly Journal of the Royal Meteorological Society*, vol. 134, no. 637, p. 1951, 2008.
- [13] G. Jiandong and Z. Gang, "Accurate estimation and application of 3-D error covariance structures in global data assimilation. Part II: background error covariance structure adjustments and numerical experiments," *Journal of Meteorological Research*, vol. 64, no. 6, pp. 669–683, 2006.
- [14] W. Jincheng, Z. Zhaorong, H. Wei, and L. Huijuan, "An improvement of background error covariance in the global GRAPES variational data assimilation and its impact on the analysis and prediction: statistics of the three-dimensional structure of background error covariance," *Journal of Meteorological Research*, vol. 72, no. 1, pp. 62–78, 2014.
- [15] M. Fisher and P. Courtier, "Estimating the covariance matrices of analysis and forecast error in variational data assimilation," *ECMWF Technical Memorandum*, vol. 220, p. 28, 1995.
- [16] M. Bonavita, L. Raynaud, and L. Isaksen, "Estimating background-error variances with the ECMWF Ensemble of Data Assimilations system: some effects of ensemble size and day-to-day variability," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 655, pp. 423–434, 2011.
- [17] Z. Weimin, C. Xiaoqun, and S. Junqiang, "Design and implementation of four-dimensional variational data assimilation system constrained by the global spectral model," *Acta Physica Sinica*, vol. 61, no. 24, 2012.
- [18] L. Bainian, Z. Weimin, C. Xiaoqun, Z. Yanlai, H. Qunbo, and L. Yu, "Investigations and experiments of variances filtering technology in the ensemble data assimilation," *The Chinese Journal of Geophysics*, vol. 58, no. 5, pp. 1526–1534, 2015.
- [19] P. Courtier, J.-N. Thépaut, and A. Hollingsworth, "A strategy for operational implementation of 4D-Var, using an incremental approach," *Quarterly Journal of the Royal Meteorological Society*, vol. 120, no. 519, pp. 1367–1387, 1994.
- [20] C. Lanczos, *Applied Analysis*, Courier Corporation, London, UK, 1988.
- [21] D. M. Barker, W. Huang, Y.-R. Guo, A. J. Bourgeois, and Q. N. Xiao, "A three-dimensional variational data assimilation system for MM5: implementation and initial results," *Monthly Weather Review*, vol. 132, no. 4, pp. 897–914, 2004.
- [22] J. Derber and F. Bouttier, "A reformulation of the background error covariance in the ECMWF global data assimilation system," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 51, no. 2, pp. 195–221, 1999.
- [23] M. Fisher, "Background error covariance modelling," in *Proceedings of the Seminar on Recent Development in Data Assimilation for Atmosphere and Ocean*, pp. 45–63, London, UK, 2003.
- [24] C. Xiaoqun, H. Sixun, Z. Weimin, and D. Huangdong, "Modeling background error covariance in regional 3D-VAR," *Science of The Total Environment*, vol. 28, no. 1, pp. 8–14, 2008.
- [25] L. Raynaud, L. Berre, and G. Desroziers, "Spatial averaging of ensemble-based background-error variances," *Quarterly Journal of the Royal Meteorological Society*, vol. 134, no. 633, 2008.