A Simultaneous Multiscale Data Assimilation Using Scale-Dependent Localization in GSI-Based Hybrid 4DEnVar for NCEP FV3-Based GFS

BO HUANG,[a] XUGUANG WANG,[a] DARYL T. KLEIST,[b] AND TING LEI[c]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/Environmental Modeling Center, College Park, Maryland*
[c] *IMSG, NOAA/Environmental Modeling Center, College Park, Maryland*

ABSTRACT: A scale-dependent localization (SDL) method was formulated and implemented in the Gridpoint Statistical Interpolation (GSI)-based four-dimensional ensemble-variational (4DEnVar) system for NCEP FV3-based Global Forecast System (GFS). SDL applies different localization to different scales of ensemble covariances, while performing a single-step simultaneous assimilation of all available observations. Two SDL variants with (SDL-Cross) and without (SDL-NoCross) considering cross-wave-band covariances were examined. The performance of two- and three-wave-band SDL experiments (W2 and W3, respectively) was evaluated through 1-month cycled data assimilation experiments. SDL improves global forecasts to 5 days over scale-invariant localization including the operationally tuned level-dependent scale-invariant localization (W1-Ope). The W3 SDL-Cross experiment shows more accurate tropical storm–track forecasts at shorter lead times than W1-Ope. Compared to the W2 SDL experiments, the W3 SDL counterparts applying tighter horizontal localization at medium-scale wave band generally show improved global forecasts below 100 hPa, but degraded global forecasts above 50 hPa. While the outperformance of the W3 SDL-NoCross experiment versus the W2 SDL-NoCross experiment below 100 hPa lasts for 5 days, that of the W3 SDL-Cross experiment versus the W2 SDL-Cross experiment lasts for 3 days. Due to local spatial averaging of ensemble covariances that may alleviate sampling error, the SDL-NoCross experiments show slightly better forecasts than the SDL-Cross experiments at shorter lead times. However, the SDL-Cross experiments outperform the SDL-NoCross experiments at longer lead times, likely from retention of more heterogeneity of ensemble covariances and resultant analyses with improved balance. Relative performance of tropical storm–track forecasts in the W2 and W3 SDL experiments are generally consistent with that of global forecasts.

KEYWORDS: Kalman filters; Variational analysis; Ensembles; Numerical weather prediction/forecasting; Data assimilation; Model initialization

1. Introduction

Ensemble-based data assimilation (DA) approaches, such as ensemble Kalman filter (EnKF; Evensen 1994) and hybrid ensemble-variational (EnVar) DA (Hamill and Snyder 2000; Lorenc 2003; Buehner 2005; Wang et al. 2007b; Wang 2010), have been widely used in many numerical weather prediction (NWP) centers to produce initial conditions for medium-range forecasts. In the ensemble-based DA approach, an ensemble of short-range forecasts is used to calculate flow-dependent background error covariances. This contrasts to the traditional three-dimensional variational DA approach that assumes static background error covariances. The advantage of the ensemble-based DA approach over the pure variational DA approach has been demonstrated in global and regional applications (Wang et al. 2007a, 2008a,b, 2013; Wang 2011; Buehner et al. 2013, 2015; Clayton et al. 2013; Gustafsson et al. 2014; Wang and Lei 2014; Lorenc et al. 2015; Kleist and Ide 2015a,b; Kutty and Wang 2015).

Limited computational resources constrain the affordable ensemble size to be much smaller than the degrees of freedom of the model itself (Houtekamer and Zhang 2016). This results in sampling error in the ensemble-based DA approach with typical features including distant spurious correlations. The successful application of an ensemble-based DA approach relies on efficient treatment of sampling error. Directly increasing ensemble size will reduce sampling error (Miyoshi et al. 2014; Lei and Whitaker 2017; Huang and Wang 2018). This, however, can be computationally prohibitive especially for operational NWP applications. Alternatively, covariance localization is applied in the ensemble-based DA approach to gradually attenuate and even eliminate distant spurious correlations caused by sampling error (Houtekamer and Mitchell 2001, 2005; Huang et al. 2019). Applying covariance localization contributes to improved analyses and subsequent forecasts (Houtekamer and Mitchell 1998, 2001; Bishop and Hodyss 2009; Buehner 2012; Anderson and Lei 2013; Gasperoni and Wang 2015).

The rapid advancement of high-performance computing allows future global NWP models to resolve a much wider range of scales. DA algorithms that appropriately update a wide range of scales will be required. Zhang et al. (2009) performed a multistep sequential DA update by separately assimilating different groups of observations and applying different localization length scales. In such a sequential approach, each group of observations by design can only update certain scales, neglecting that observations contain information that can be used to update all resolved scales (Caron and Buehner 2018).

Corresponding author: Xuguang Wang, xuguang.wang@ou.edu

Miyoshi and Kondo (2013) combined two sets of independent analysis increments from assimilating the same set of observations, with each set utilizing different amounts of localization. Both methods showed improved analyses and subsequent forecasts in the EnKF systems compared to applying fixed uniform localization once at all scales.

While the aforementioned methods took multiple steps or adopted a sequential update, a single-step simultaneous multiscale update was proposed recently by introducing scale-dependent localization (SDL) in the EnVar framework (Buehner 2012; Buehner and Shlyaeva 2015) and in the pure EnKF framework (Wang et al. 2021). The simultaneous SDL in EnVar can be classified into two variants. The first variant completely eliminates the cross-wave-band covariances (Buehner 2012) (hereafter, referred to as SDL-NoCross). Mathematically, SDL-NoCross equivalently applies a local spatial averaging of ensemble covariances, which may alleviate sampling error and improve the accuracy of ensemble covariances especially for a small ensemble (Buehner and Charron 2007). Compared to applying a fixed uniform localization once at all scales, SDL-NoCross improved the general global forecast skill in a global 3DEnVar (Buehner 2012) and 4DEnVar (Lorenc 2017) system.

The second simultaneous SDL variant takes into account cross-wave-band covariances (hereafter, referred to as SDL-Cross) (Buehner and Shlyaeva 2015). Compared to SDL-NoCross, SDL-Cross may retain more heterogeneity of error covariances (Caron and Buehner 2018). SDL-Cross was demonstrated in a regional 3DEnVar sea ice DA system to perform better than fixed uniform localization at all scales (Buehner and Shlyaeva 2015). Caron and Buehner (2018) implemented SDL-Cross in a global 3DEnVar system, and found improved global forecasts over scale-invariant localization. Furthermore, Caron et al. (2019) comparing SDL-NoCross and SDL-Cross in a regional 3DEnVar system showed that SDL-NoCross produced more accurate forecasts than SDL-Cross using a 25-member ensemble, while both performed comparably when using a 75-member ensemble populated by time-lagged method (Van Den Dool and Rukhovets 1994; Lorenc 2017; Huang and Wang 2018). Caron et al. (2019) further hypothesized that the relative performance between SDL-NoCross and SDL-Cross could be associated with the accuracy of the estimated cross-wave-band covariances in SDL-Cross that depended on ensemble size.

This paper addresses several additional questions on simultaneous SDL using the U.S. NWS GSI-based hybrid 4DEnVar system (Wang and Lei 2014; Kleist and Ide 2015a). The GSI-based 4DEnVar system was recently integrated with the U.S. next-generation nonhydrostatic Finite-Volume Cubed-sphere dynamical core (FV3)-based GFS model (JCSDA 2018; Chen et al. 2019; Zhou et al. 2019). To achieve the goal of exploring new scientific questions associated with simultaneous SDL, we first demonstrated both SDL methods mathematically in an EnVar framework that is preconditioned on the full background error covariances (hereafter, referred to as **B**; Derber and Rosati 1989; Wang 2010), and implemented both in the GSI-based hybrid 4DEnVar system. SDL formulations within an EnVar framework preconditioned on the full **B** and square root of **B** were also discussed in Caron et al. (2019).

Given that SDL-NoCross and SDL-Cross were only compared for regional applications previously, this paper first evaluates and compares both approaches for global forecasts. Second, in previous studies, SDL was implemented with no explicit level dependence and compared with level- and scale-invariant horizontal localization. In our study, the baseline operational GSI-based hybrid 4DEnVar system for the FV3-based GFS applies scale-invariant, but level-dependent horizontal localization. Therefore, how does SDL perform relative to the stricter baseline that adopts scale-invariant, but level-dependent horizontal localization? Third, in a global modeling system, a tropical storm is at relatively small scale and strongly influenced by the general large-scale environmental flow (Zong and Wu 2015). How would SDL-NoCross and SDL-Cross influence the tropical storm–track forecasts in the FV3-based GFS? Fourth, efficient scale separation in SDL is essential and remains to be investigated. How does the performance of SDL vary with different numbers of decomposed wave bands (e.g., two versus three)? Finally, diagnostics were performed to understand the different performance between scale-invariant localization, SDL-NoCross and SDL-Cross.

This paper is organized as follows. Section 2 describes the SDL formulation and implementation in the GSI-based hybrid 4DEnVar system. Experiment design is described in section 3. Sections 4 and 5 discuss the experiment results. In section 6, some diagnostics are presented to understand the results in sections 4 and 5. The computational cost is compared in section 7. Section 8 presents the conclusions and discussion.

2. SDL formulation and implementation in the GSI-based 4DEnVar system

a. General SDL formulation in the GSI-based 4DEnVar system

The GSI-based 4DEnVar system is formulated and implemented based on the extended control variable method, to incorporate the ensemble background covariances within the traditional variational framework (Wang 2010; Wang et al. 2013; Wang and Lei 2014). Mathematically, it is equivalent to linearly combining the static and ensemble background covariances (Wang et al. 2007b, 2008a). In this study, SDL was implemented in the GSI-based hybrid 4DEnVar system by further extending control variables. In this section, the general SDL formulation in the 4DEnVar system is first illustrated following the notations in Wang et al. (2013) and Wang and Lei (2014). Specific implementations of SDL-NoCross and SDL-Cross are then detailed. To highlight the variables of further extended dimension due to applying SDL, a "hat" sign is labeled above the letter or symbol.

Following the study of Buehner and Shlyaeva (2015), the ensemble perturbations are first decomposed into J overlapping wave bands or scales indexed by $j = 1, \ldots, J$:

$$\mathbf{x}_{k,j}^e = \mathbf{\Psi}_j \mathbf{x}_k^e, \tag{1}$$

where $\mathbf{x}_k^e$ denotes the kth ensemble perturbation vector normalized by $(K-1)^{1/2}$, K is the ensemble size, $\mathbf{\Psi}_j$ is the spectral

filter function that extracts the jth wave band, and $\mathbf{x}^e_{k,j}$ is the kth normalized ensemble perturbation vector that only contains the jth wave band.

In 4DEnVar with SDL, the analysis increment $\mathbf{x}'_t$ at time t in a DA window is calculated as

$$\mathbf{x}'_t = \mathbf{x}'_1 + \hat{\mathbf{I}} \sum_{k=1}^{K} [\hat{\mathbf{a}}_k \circ (\hat{\mathbf{x}}^e_k)_t], \qquad (2)$$

where

$$\hat{\mathbf{I}} = [\mathbf{I}\ \mathbf{I}\ \cdots \ \mathbf{I}], \qquad (3)$$

$$\hat{\mathbf{a}}_k = \begin{bmatrix} \mathbf{a}_{k,1} \\ \mathbf{a}_{k,2} \\ \vdots \\ \mathbf{a}_{k,J} \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{x}}^e_k = \begin{bmatrix} \mathbf{x}^e_{k,1} \\ \mathbf{x}^e_{k,2} \\ \vdots \\ \mathbf{x}^e_{k,J} \end{bmatrix}. \qquad (4)$$

The first term on the right-hand side of Eq. (2) is the analysis increment associated with the static background error covariances. Matrix $\hat{\mathbf{I}}$ contains J identity matrices aligned in a row. $(\hat{\mathbf{x}}^e_k)_t$ is a vector that concatenates J vectors of decomposed $\mathbf{x}^e_{k,j}$ $(j = 1, \ldots, J)$ for the kth member at time t. The three-dimensional vector $\mathbf{a}_{k,j}$ corresponds to the control variable vector at the jth wave band for the kth member, and $\hat{\mathbf{a}}_k$ is further extended control variable vector that concatenates J vectors of $\mathbf{a}_{k,j}$ for the kth member. The sign $\circ$ denotes the Schur product. Compared to Wang and Lei (2014) without applying SDL, the control variable vector $\hat{\mathbf{a}}_k$ in Eq. (4) for the kth member varies with the wave-band index j. Its dimension is increased by a factor of J. As in Wang and Lei (2014), the same set of $\hat{\mathbf{a}}_k$ is applied for all control variables that include surface pressure, virtual temperature, wind, relative humidity, cloud water, and ozone mixing ratio at different time levels.

The analysis increment in Eq. (2) is obtained by minimizing the following cost function:

$$J(\mathbf{x}'_1, \hat{\mathbf{a}}) = \frac{1}{2}\beta_1 (\mathbf{x}'_1)^\mathrm{T} \mathbf{B}_1^{-1}(\mathbf{x}'_1) + \frac{1}{2}\beta_2 (\hat{\mathbf{a}})^\mathrm{T} \hat{\mathbf{A}}^{-1}(\hat{\mathbf{a}})$$
$$+ \frac{1}{2}\sum_{t=1}^{L} (\mathbf{y}^{o\prime}_t - \mathbf{H}_t \mathbf{x}'_t)^\mathrm{T} \mathbf{R}_t^{-1}(\mathbf{y}^{o\prime}_t - \mathbf{H}_t \mathbf{x}'_t). \qquad (5)$$

On the right-hand side of Eq. (5), the first term is associated with the static background covariance $\mathbf{B}_1$. In the second term, $\hat{\mathbf{a}}$ is the extended control variable vector that concatenates K vectors of $\hat{\mathbf{a}}_k$ in Eq. (4). The block-diagonal matrix $\hat{\mathbf{A}}$ defines the localization matrix for within- and cross-wave-band ensemble covariances (see more details later). In the third term, $\mathbf{y}^{o\prime}_t$, $\mathbf{H}_t$, and $\mathbf{R}_t$ are the observation innovation vector, linearized observation operator matrix, and observation covariance matrix at time t, respectively. The upper bound of summation L is the number of time levels spanning the DA window (e.g., 6 h). In addition, parameters β_1 and β_2 control the weights of the static and ensemble background covariances. In the operational global GSI-based 4DEnVar system at NCEP, $(1/\beta_1) + (1/\beta_2) = 1$ is defined to maintain the
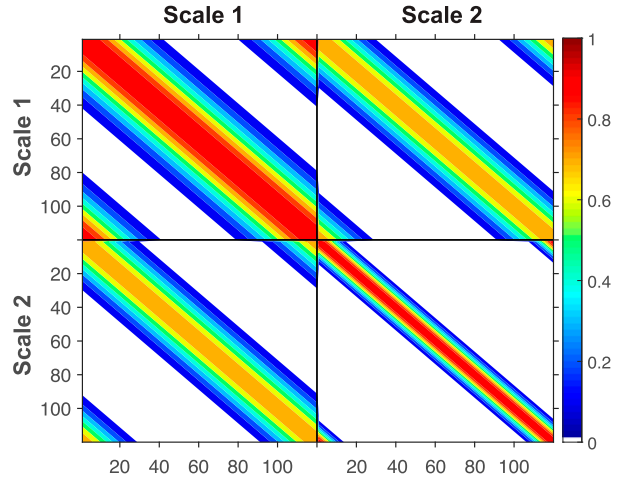

FIG. 1. Illustration of complete scale-dependent spatial localization matrix between Scale 1 and Scale 2 that represent large and small scales, respectively, using a one-dimensional periodic domain of 100 grid points.

total background variances following Wang et al. (2007b, 2008a) and Wang (2010).

In the global GSI-based 4DEnVar system, the localization defined in $\hat{\mathbf{A}}$ is realized through spectral filter transformation in the horizontal direction and recursive filter transformation in the vertical direction. More details of implementing the horizontal and vertical localization were described in Wang (2010) and Wang et al. (2013). In this study, SDL is only applied for horizontal localization. Specifically, the explicit formula of $\hat{\mathbf{A}}$ can be written as

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{\mathbf{L}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{L}} \end{bmatrix}. \qquad (6)$$

Each of the K blocks in $\hat{\mathbf{A}}$ contains the same predefined localization matrix $\hat{\mathbf{L}}$ with unit diagonal elements following Buehner and Shlyaeva (2015):

$$\hat{\mathbf{L}} = \begin{bmatrix} \hat{\mathbf{L}}_{1,1} & \hat{\mathbf{L}}_{1,2} & \cdots & \hat{\mathbf{L}}_{1,J} \\ \hat{\mathbf{L}}_{2,1} & \hat{\mathbf{L}}_{2,2} & \cdots & \hat{\mathbf{L}}_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{L}}_{J,1} & \hat{\mathbf{L}}_{J,2} & \cdots & \hat{\mathbf{L}}_{J,J} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{L}}^{1/2}_{1,1} \\ \hat{\mathbf{L}}^{1/2}_{2,2} \\ \vdots \\ \hat{\mathbf{L}}^{1/2}_{J,J} \end{bmatrix}$$
$$\times \begin{bmatrix} \hat{\mathbf{L}}^{\mathrm{T}/2}_{1,1} & \hat{\mathbf{L}}^{\mathrm{T}/2}_{2,2} & \cdots & \hat{\mathbf{L}}^{\mathrm{T}/2}_{J,J} \end{bmatrix}, \qquad (7)$$

where $\hat{\mathbf{L}}_{j1,j2} = \hat{\mathbf{L}}^{1/2}_{j1,j1} \hat{\mathbf{L}}^{\mathrm{T}/2}_{j2,j2}$ $(j1 = 1, \ldots, J$ and $j2 = 1, \ldots, J)$ defines the localization matrix for the ensemble covariances between the $j1$th and $j2$th wave bands. Equation (7) ensures the complete localization matrix is positive semidefinite (Buehner and Shlyaeva 2015). Given the design in Eq. (7), the within-wave-band localization matrices have unit diagonal elements, while the cross-wave-band localization matrices display less-than-one diagonal elements (Fig. 1).

TABLE 1. List of DA experiments.

Expt	No. of wave bands	Horizontal localization length scale (*e*-folding distance)	Vertical localization length scale (scale height, *e*-folding distance)
W1-Ope	1	Level-dependent length scale for full-scale ensemble perturbations (black curve in Fig. 2)	0.5
W1-1000	1	1000 km for full-scale ensemble perturbations	
W1-650	1	650 km for full-scale ensemble perturbations	
W1-300	1	300 km for full-scale ensemble perturbations	
W2-NoCross	2	1000 and 300 km for large- and small-scale ensemble perturbations,	
W2-Cross	2	respectively	
W3-NoCross	3	1000, 650, and 300 km for large-, medium-, and small-scale	
W3-Cross	3	ensemble perturbations, respectively	

b. *Specific implementations of SDL-NoCross and SDL-Cross in the GSI-based 4DEnVar system*

In this subsection, implementations of SDL-NoCross and SDL-Cross in the GSI-based 4DEnVar system are further described. Two major implementation differences are involved. One is the definition of the spectral filter function $\mathbf{\Psi}_j$ in Eq. (1). The other is whether or not to zero out the cross-wave-band localization matrix $\hat{\mathbf{L}}_{j1,j2}(j1 \neq j2)$ in Eq. (7).

In SDL-Cross, it is required that the spectral filter functions $\mathbf{\Psi}_j$ over J wave bands sum to one to recover the original raw ensemble perturbations from their decomposed components (Buehner and Shlyaeva 2015), and that the cross-wave-band localization matrix $\hat{\mathbf{L}}_{j1,j2}$ ($j1 \neq j2$) is retained to partially include cross-wave-band covariances. In a particular scenario of applying the same amount of localization at different wave bands, SDL-Cross is equivalent to applying fixed uniform localization once at all scales (Buehner and Shlyaeva 2015).

In contrast, SDL-NoCross requires that the squared spectral filter functions $\mathbf{\Psi}_j$ over J wave bands sum to one to maintain the total raw ensemble variances (Buehner 2012), and that the cross-wave-band localization matrix $\hat{\mathbf{L}}_{j1,j2}$ ($j1 \neq j2$) is set to be zero to completely remove cross-wave-band covariances.

3. Experiment design

The GSI-based 4DEnVar DA system and the FV3-based GFS model were used for 1-month cycled DA experiments from 0000 UTC 25 August to 1800 UTC 24 September 2017. More details about the GSI-based EnVar system can be found in Wang et al. (2013) and Wang and Lei (2014). A general flowchart of the GSI-based EnVar system was shown in Fig. 1b of Wang et al. (2013) that hybridizes the EnVar and EnKF components for one DA cycle. The assimilated observations over a 6-h DA window include all the conventional and satellite observations in the operational NCEP global DA system (http://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/table_2.htm and table_18.htm). Satellite radiance data assimilation applied the same observation quality control and bias correction in the operational global DA system (Zhu et al. 2014).

The baseline 4DEnVar experiment (W1-Ope in Table 1) was set up similarly to the operational system using a dual-resolution configuration (JCSDA 2018). A one-member control

background is at a resolution of C384 (~25 km), while an 80-member ensemble background is at a reduced resolution of C192 (~50 km) (Putman and Lin 2007). In the EnVar update, the one-member control background was updated using the 4DEnVar algorithm (Wang and Lei 2014; Kleist and Ide 2015a). Specifically, 12.5% static and 87.5% ensemble background covariances were combined to construct its hybrid form as in the operational system. The 3-hourly ensemble perturbations[1] were ingested in the 4DEnVar update to account for the temporal evolution. To deal with sampling error, localization was applied in both horizontal and vertical directions. As in the operational system, the horizontal localization length scale varies with model level (e.g., black dotted curve in Fig. 2), and a fixed uniform vertical localization length scale is applied at all model levels (Table 1). In addition, a tangent linear normal mode initialization constraint (TLNMC, Kleist et al. 2009) was applied to alleviate the imbalance in the control analysis as in the operational system.

In the EnKF update, the 4D local ensemble transform Kalman filter (LETKF, Bishop et al. 2001; Hunt et al. 2007) was adopted to update the 80-member background ensemble. In the LETKF, the observation operators were calculated through the GSI. To remedy sampling error, localization was applied in the LETKF and defined by the Gaspari–Cohn function (Gaspari and Cohn 1999). The exact cutoff distance in the Gaspari–Cohn function is equivalent to multiplying the *e*-folding localization length scales in the 4DEnVar update by a factor of 2.577 (Pan et al. 2014). To remedy the background ensemble spread deficiency, multiplicative inflation (Whitaker and Hamill 2012) was employed to relax the posterior ensemble spread back to 85% of the prior ensemble spread. Stochastic parameterization schemes (Palmer et al. 2009; Lei and Whitaker 2016, 2017; Huang and Wang 2018) were further applied to account for model uncertainty.

The GFS model using the nonhydrostatic Finite-Volume cubed-sphere dynamical core (FV3; Lin 2004; Harris and Lin 2013) was used to provide the control and ensemble background forecasts. The FV3-based GFS was configured similarly as in the preoperational tests in Phase II of the Next-Generation Global

---

[1] Operational 4DEnVar used 1-hourly interval. We adopt 3-hourly interval for this study to save computational cost.
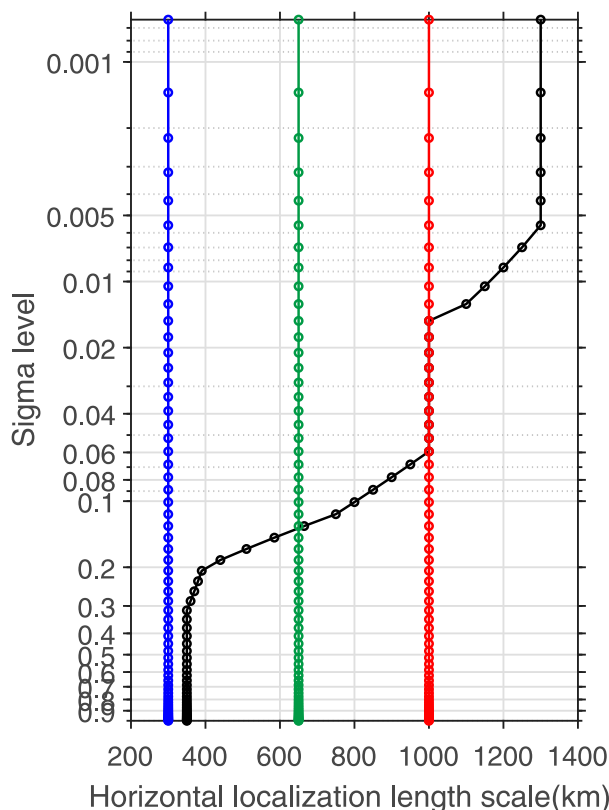
FIG. 2. Level-dependent horizontal localization length scales (unit: km) for W1-Ope (black) and level-invariant horizontal localization length scales at large- (red), medium- (green), and small- (blue) scale wave bands in the two- and three-wave-band SDL experiments in Table 1. Note that the horizontal localization length scale is *e*-folding distance.

Prediction System (NGGPS) project (https://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics). The model configurations were detailed in Chen et al. (2019) and Zhou et al. (2019). Due to computational constraints, the experiments in this study were performed at a reduced horizontal resolution compared to the preoperational tests. There are a total of 64 model levels in the current FV3-based GFS model. The 4D incremental analysis update (4DIAU) was further applied for both control and ensemble forecasts to improve the balance during the model integration (Bloom et al. 1996; Lorenc et al. 2015; Lei and Whitaker 2016, 2017; Huang and Wang 2018).

In SDL, the optimal way of performing scale separation remains to be investigated. A tropical cyclone example (Fig. 3a) was used to assist in scale separation in our SDL experiments. As our initial examination, the first set of SDL experiments adopted two wave bands, referred to as W2-NoCross and W2-Cross in Table 1 that apply SDL-NoCross and SDL-Cross, respectively. For the scale separation in the two-wave-band SDL experiments, the tropical cyclone at relatively small scale (Fig. 3c) was isolated from the general large-scale environmental flow (Fig. 3b) using the example in Fig. 3a. The spectral decomposition was performed using spherical harmonic

transform built in the operational GSI-based 4DEnVar system. The resultant spectral filter functions (Figs. 4a,b) at the large- and small-scale wave bands cross each other roughly at wavelength 2500 km or at total wavenumber 16. The spectral filter functions in SDL-NoCross were defined as the square root of those in SDL-Cross following Caron et al. (2019). It is meant to satisfy that the spectral filer functions sum to one in SDL-Cross, while the squared spectral filter functions sum to one in SDL-NoCross as discussed in section 2b. To define the horizontal localization length scale in the two-wave-band SDL experiments, several tuning tests were performed by referring to the horizontal localization length scales applied in Caron and Buehner (2018) and the operational level-dependent horizontal localization length scales in W1-Ope. Finally, 1000 and 300 km *e*-folding distances were selected as the horizontal localization length scales at all model levels for the large- and small-scale ensemble perturbations in the two-wave-band SDL experiments, respectively. To provide a more homogeneous comparison with the two-wave-band SDL experiments applying level-invariant horizontal localization, another three experiments of W1-1000, W1-650, and W1-300 (Table 1) were designed. Different from W1-Ope applying the level-dependent horizontal localization, however, W1-1000, W1-650, and W1-300 apply the level-invariant horizontal localization length scales, that is, 1000, 650, and 300 km *e*-folding distances, respectively. Due to limited computational resources, only some simple diagnostics in sections 4a and 6a, rather than the full 1-month cycled DA experiments, were performed in W1-650 to investigate if SDL is equivalent to applying a fixed intermediate localization length scale.

To explore how the SDL performance varies with the number of decomposed wave bands, three-wave-band SDL experiments were further designed such as W3-NoCross and W3-Cross in Table 1 that apply SDL-NoCross and SDL-Cross, respectively. Given the amount of experiments to be conducted and the constraint of computational resources, only two and three wave bands are experimented. For the scale separation in the three-wave-band SDL experiments, the large-scale background in the two-wave-band SDL experiments was further decomposed to two wave bands, that is, the large- and medium-scale wave bands in the three-wave-band SDL experiments. As a result, the midlatitude high pressure (Fig. 3e) that appears in the large-scale wave band in the two-wave-band SDL experiments is further isolated from the global large-scale environmental flow (Fig. 3d). The background of the small-scale wave band in the three-wave-band SDL experiments (Fig. 3f) was retained similarly as in the two-wave-band SDL experiments (Fig. 3c) typically featured by the tropical cyclone. Figures 4c and 4d show the resultant spectral filter functions at large-, medium-, and small-scale wave bands in the three-wave-band SDL experiments. Additional tuning tests were performed to determine the horizontal localization length scales in the three-wave-band SDL experiments. It did not provide additional benefits by further increasing the horizontal localization length scale beyond 1000 km *e*-folding distance for the large-scale wave band in the three-wave-band SDL experiments. The same set of 1000 and 300 km *e*-folding distances were thus applied at the large- and small-scale wave bands in
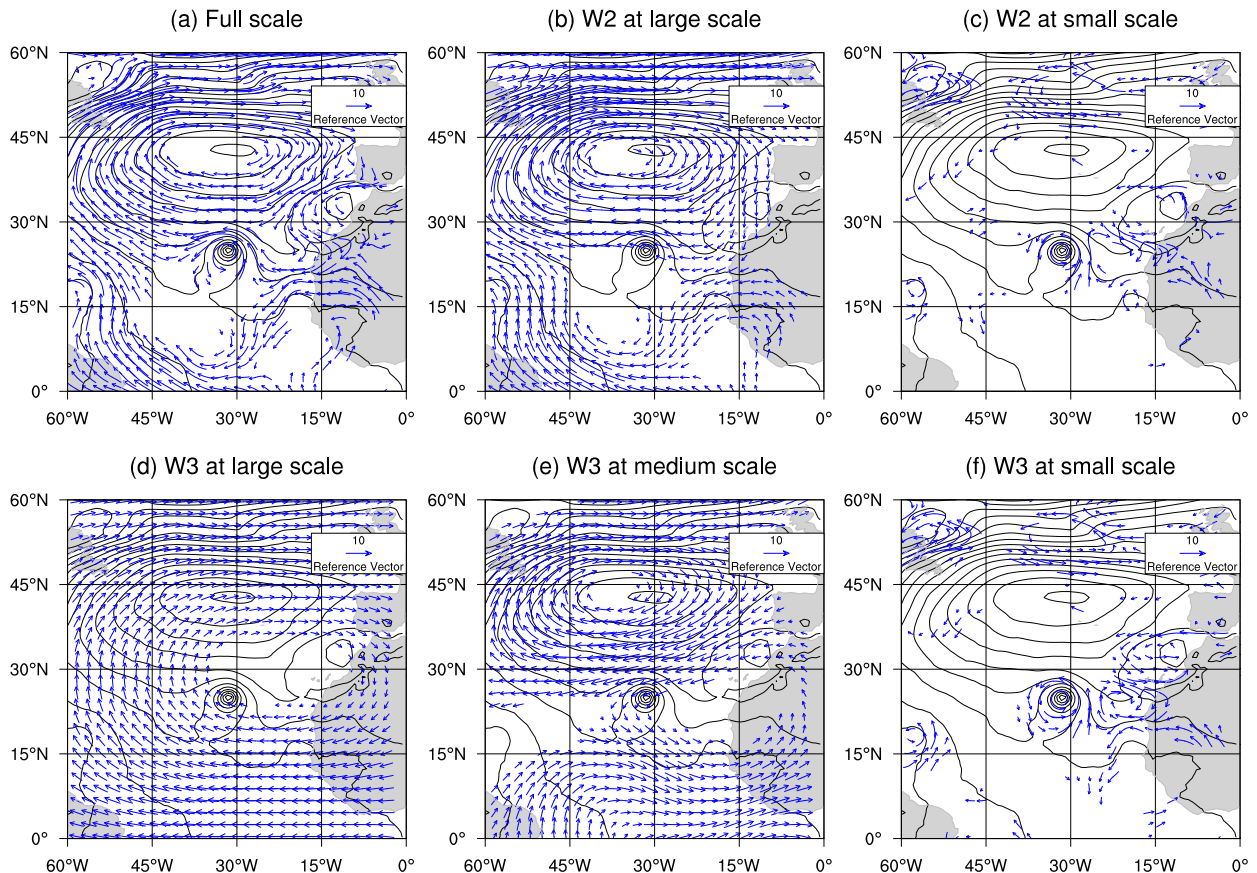
Fig. 3. Two-dimensional wind vectors (unit: m s^{-1}) at 850 hPa in a tropical cyclone example at 1800 UTC 13 Sep 2013 at (a) full-scale, and at (b) large- and (c) small-scale wave bands in the two-wave-band SDL experiments, and at (d) large-, (e) medium-, and (f) small-scale wave bands in the three-wave-band SDL experiments. The gray contours beneath the wind vectors denote the geopotential height from 1420 to 1640 gpm with an interval of 20 gpm at 850 hPa at 1800 UTC 13 Sep 2013.

the three-wave-band SDL experiments. Additionally, a 650 km e-folding distance was selected at the medium-scale wave band in the three-wave-band SDL experiments. Therefore, the main difference in the two- and three-wave-band SDL experiments is tighter horizontal localization applied at medium-scale wave band in the three-wave-band SDL experiments. This would facilitate interpreting performance differences between the two- and three-wave-band SDL experiments.

Experiment descriptions are detailed in Table 1. For all of the experiments in Table 1, they applied the 0.5 scale-height e-folding distance at all model levels for the vertical localization. On the other hand, in the LETKF update, all of the experiments applied the same horizontal and vertical localization length scales as in W1-Ope.

To provide a robust comparison among different experiments, a paired permutation test was applied at 95% confidence level with 1000 replicates (Manly 2006). Procedures in Wang and Bishop (2005) were followed to collect independent samples for the significance test. For the global forecast comparison in sections 4c, 4d and 5a, the time series were first averaged in each of the independent subdomains distributed over the globe. This produced spatially independent time series. For

the TC track forecast comparison in sections 4e and 5b, each storm was treated as an independent case. For the power spectral evaluation in section 6c, each model level was treated as spatially independent. Then, for each set of spatially independent time series, lag correlations were computed to further determine the length of a temporal block so that the temporal block series were weakly correlated in time. The objective of the aforementioned procedures was to determine effective degrees of freedom during the significance test. All sets of the resultant temporally and spatially independent sample time series were pooled together to perform the paired permutation test. A false discovery rate method (Wilks 2006) was further applied to ameliorate the simultaneous multiple hypothesis test issue.

4. Comparison of two-wave-band SDL and scale-invariant localization experiments

a. Single observation experiment

To demonstrate the impacts of SDL, a single observation experiment was first performed for W1-1000, W1-650, W1-300,
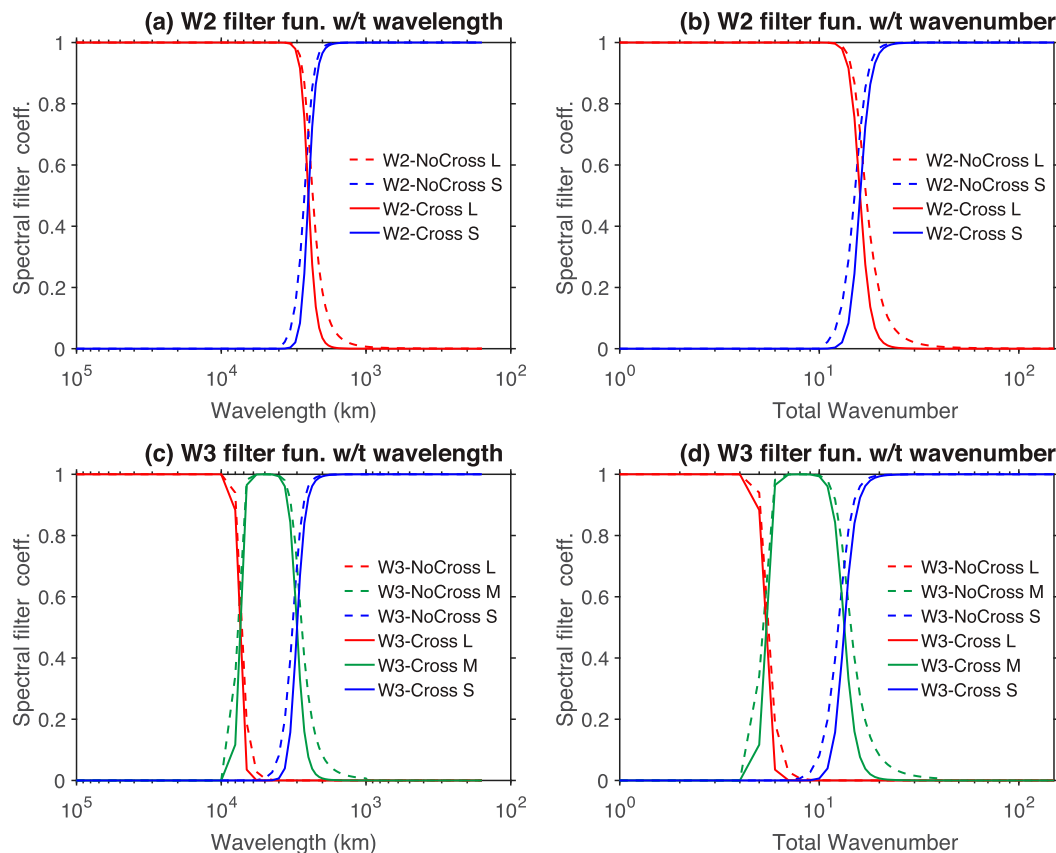
FIG. 4. Spectral filter functions for SDL-NoCross (dashed) and SDL-Cross (solid) in the (a),(b) two- and (c),(d) three-wave-band SDL experiments at large- (L, red), medium- (M, green) and small- (S, blue) scale wave bands as a function of the (a),(c) wavelength and (b),(d) total wavenumber. Note that the two-wave-band SDL experiments in (a) and (b) only contain the large- and small-scale wave bands.

W2-NoCross, and W2-Cross (Fig. 5). The case that likely featured an interaction between a large-scale subtropical high and mesoscale tropical cyclone was selected to more clearly reveal the differences between SDL-Cross and SDL-NoCross. As expected, in the scale-invariant localization experiments, a tighter horizontal localization length scale produces a more localized analysis increment pattern. In particular, in contrast to W1-300 applying the tightest horizontal localization length scale (Fig. 5c), W1-1000 and W1-650 applying relatively wider horizontal localization length scales produce two analysis increment maxima that are located at the observation location and to the north of the tropical cyclone center (Figs. 5a,b), respectively. The distant analysis increment maximum to the north of the tropical cyclone center may suggest that the raw ensemble covariances estimate an interaction between the subtropical high and tropical cyclone. In addition, by applying the same horizontal localization length scale 1000 km e-folding distance for both large- and small-scale wave bands in W2-Cross, it reproduces the same analysis increment pattern as W1-1000 (not shown here). This is consistent with the theory that applying the same amount of localization at different wave bands in SDL-Cross is equivalent to applying fixed uniform localization once at all scales (Buehner and Shlyaeva 2015).

This further suggests that SDL was correctly implemented in the GSI-based 4DEnVar system. Due to applying a wider horizontal localization length scale for large-scale ensemble perturbations, W2-NoCross and W2-Cross show broader analysis increment patterns than W1-300. Compared to W1-1000, W2-NoCross and W2-Cross show overall more restricted analysis increment patterns due to much tighter horizontal localization applied for small-scale ensemble perturbations. In addition, different analysis increment patterns between W1-650 and W2-NoCross/W2-Cross suggest that SDL is not equivalent to applying a fixed intermediate localization length scale.

In addition, W2-NoCross only shows one analysis increment maximum at the observation location, while W2-Cross that partially includes cross-wave-band covariances maintains two analysis increment maxima as in W1-1000 and W1-650. This result suggests that the distant analysis increment maximum to the north of the tropical cyclone center in W1-1000, W1-650, and W2-Cross is contributed by the cross-wave-band covariances between the subtropical high and tropical cyclone. The reduced magnitude of the analysis increment maxima to the north of the tropical cyclone center in W2-Cross relative to W1-1000 and W1-650 can be partially attributed to the
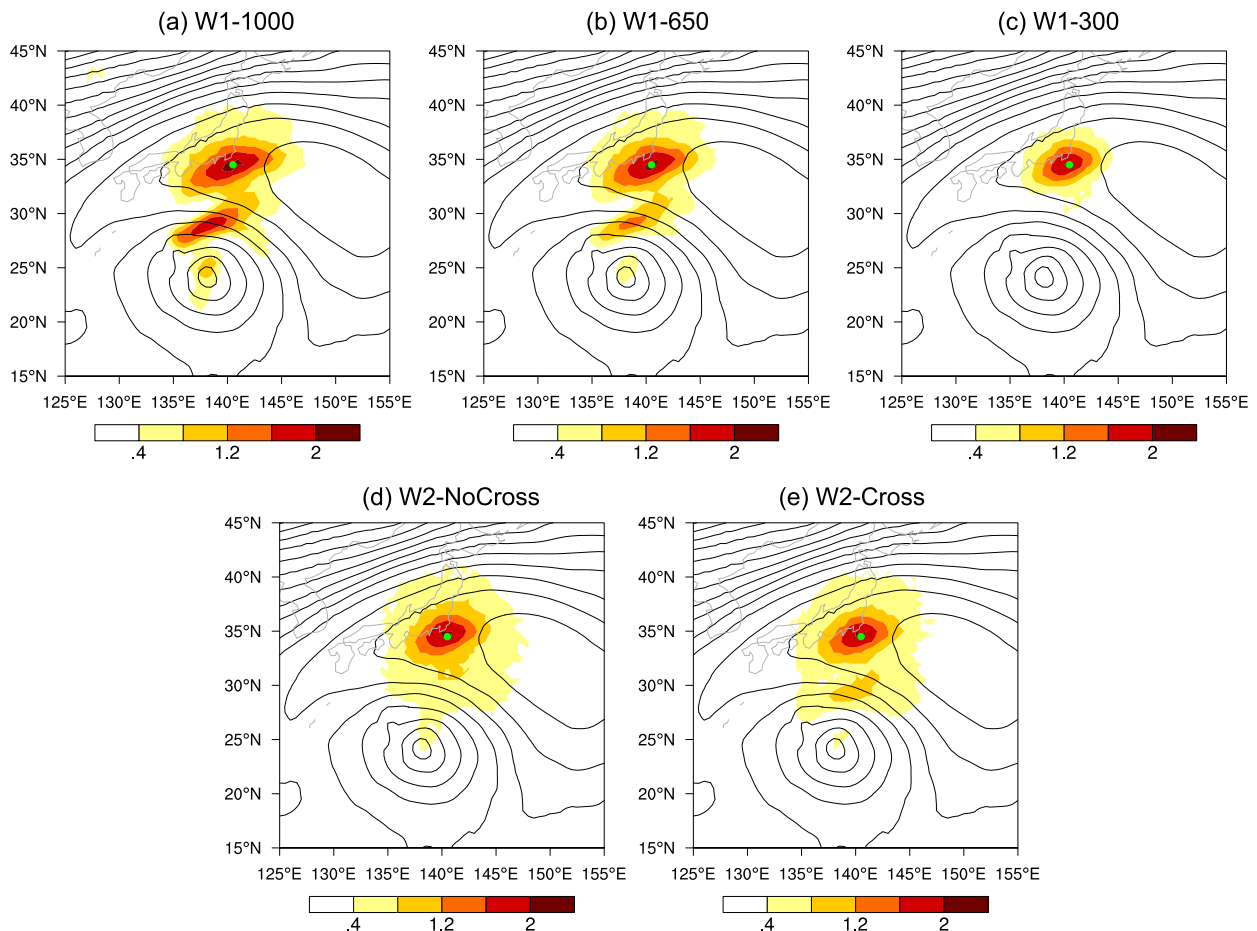
(a) W1-1000

(b) W1-650

(c) W1-300

(d) W2-NoCross

(e) W2-Cross

FIG. 5. 500 hPa meridional wind analysis increments (unit: m s^{-1}) in a tropical cyclone example at 1800 UTC 13 Sep 2013 from assimilating a single meridional wind observation at the green dot that is 5 m s^{-1} higher than the background in (a) W1-1000, (b) W1-650, (c) W1-300, (d) W2-NoCross, and (e) W2-Cross. The gray contours denote the full-scale geopotential height from 5760 to 5900 gpm with an interval of 20 gpm at 500 hPa at 1800 UTC 13 Sep 2013.

less-than-one peak in the cross-wave-band localization matrix applied in SDL as shown in Fig. 1. As expected, a further inspection of decomposed large- and small-scale analysis increments shows that W2-NoCross and W2-Cross show similar large-scale analysis increment patterns as W1-1000 and similar small-scale analysis increment patterns as W1-300 (not shown).

b. Analysis increment power

To investigate how the localization influences the analysis increments at different scales, the analysis increment power spectrum was calculated as a function of total wavenumber and averaged over the cycled DA period. Figure 6 shows the averaged analysis increment power spectra for the temperature and wind variables at 500 hPa. By applying a much wider horizontal localization length scale at 500 hPa (roughly at the 0.5 sigma level in Fig. 2), W1-1000 shows larger analysis increment power at all total wavenumbers compared to W1-300 and W1-Ope. W1-300 shows slightly reduced analysis increment power than W1-Ope, since it applies slightly tighter horizontal localization at 500 hPa compared to W1-Ope. These

results suggest that broader horizontal localization generally produces larger analysis increment power in our cycled DA experiments. Furthermore, the analysis increment power in W2-NoCross and W2-Cross is at a magnitude closer to W1-1000 at small total wavenumbers and closer to W1-300 at large total wavenumbers. This is to be expected because the two-wave-band SDL experiments apply the same horizontal localization length scales as W1-1000 and W1-300 at small and large total wavenumbers, respectively. The analysis increment power differences between W2-NoCross and W2-Cross are slightly more noticeable at small total wavenumbers.

c. Forecast verification against rawinsonde observations

Root-mean-square-errors (RMSEs) were calculated by comparing the 6-h background forecasts against rawinsonde observations (Fig. 7). In both subsections 4c and 4d, W1-Ope is used as a reference for comparison. As discussed in section 3, W1-Ope adopts the operationally tuned, level-dependent, scale-invariant localization, and therefore provides a stricter reference compared to the level and scale-invariant localization
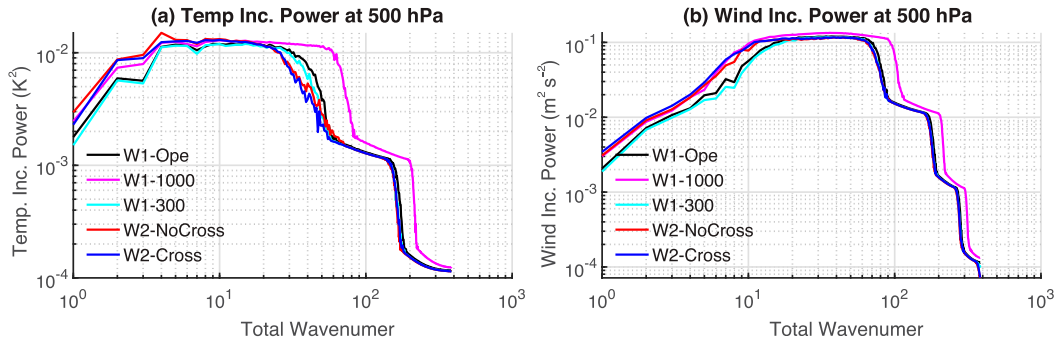
FIG. 6. Averaged analysis increment power spectra over the cycled DA period as a function of total wavenumber for the (a) temperature (unit: K^2) and (b) wind (unit: $m^2 s^{-2}$) variables at 500 hPa in W1-Ope (black), W1-1000 (magenta), W1-300 (cyan), W2-NoCross (red), and W2-Cross (blue).

experiments, W1-300 and W1-1000. By applying a much wider horizontal localization length scale for the full-scale ensemble perturbations, W1-1000 shows the worst 6-h temperature and wind forecasts for most vertical levels. Compared to W1-Ope, W1-300 shows comparable or slightly improved 6-h background forecasts at several levels below 100 hPa especially for the wind forecasts. This suggests that slightly reducing the horizontal localization length scale below 100 hPa in W1-Ope may further benefit the forecasts at such levels. However, W1-300 produces larger temperature forecast error than W1-Ope above 100 hPa. It is also noticed that W1-1000 has the least degradation relative to W1-Ope above 100 hPa compared to lower levels. These results together suggest that a wider horizontal localization length scale for the full-scale ensemble perturbations is generally beneficial at upper model levels. This is consistent with the current operational level-dependent horizontal localization settings in W1-Ope, and may explain

the degradation in W1-300 above 100 hPa that applies much tighter horizontal localization.

On the other hand, W2-NoCross and W2-Cross generally improve the 6-h global forecasts over W1-Ope, W1-1000 and W1-300 at most levels, suggesting the benefits of SDL. In particular, W2-NoCross and W2-Cross that apply tighter horizontal localization for the small-scale ensemble perturbations show the largest improvement for 6-h temperature and wind forecasts over W1-Ope above 200 hPa. Further diagnostic shows that the percentage of variance projected onto large scales increases above 200 hPa (not shown). The largest improvement above 200 hPa in W2-NoCross and W2-Cross is likely to be related to more contributions from the wider horizontal localization length scale applied at large-scale wave band in the two-wave-band SDL experiments. The relative performance between W2-NoCross and W2-Cross is a bit mixed for the 6-h background forecasts below 300 hPa.
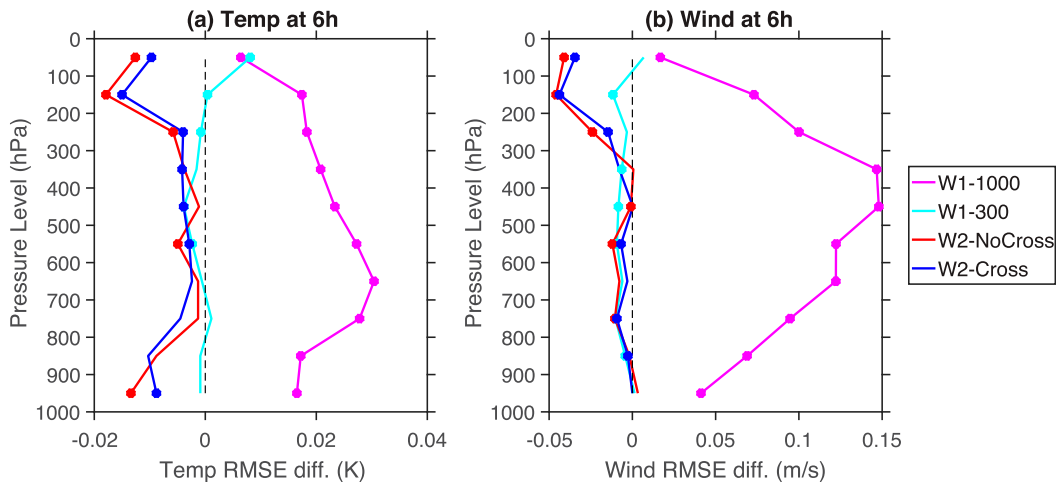

FIG. 7. Difference of the root-mean-square-error (RMSE) of the 6-h (a) temperature (unit: K) and (b) wind background (unit: $m s^{-1}$) forecasts against the rawinsondes as a function of pressure level in W1-1000 (magenta), W1-300 (cyan), W2-NoCross (red), and W2-Cross (blue) relative to W1-Ope. Negative (positive) values mean improved (degraded) 6-h background forecasts relative to W1-Ope. The bold dot sign indicates the RMSE difference is statistically significant by applying a paired permutation test with the false discovery method at 5% significance level. The dashed black line denotes zero values.

However, W2-NoCross slightly improves the 6-h background forecasts more than W2-Cross above 300 hPa. This result may suggest the cross-wave-band covariances estimated by the ensemble may not be reliable above 300 hPa, consistent with the hypothesis in Caron et al. (2019) that relative performances between SDL-NoCross and SDL-Cross depend on the accuracy of estimated cross-wave-band covariances.

d. Global forecast verification against ECMWF reanalyses (ERA-Interim)

To evaluate medium-range global forecasts, root-mean-square-errors (RMSEs) between the 5-day global forecasts and ECMWF interim reanalysis (ERA-Interim; Dee et al. 2011) were calculated at selected levels every 6 h. Figure 8 shows the RMSE difference relative to W1-Ope as a function of forecast lead time and pressure level. W1-1000 in general shows the worst 5-day global temperature and wind forecasts than W1-Ope. This further suggests negative impacts of applying much wider horizontal localization lengths scale at these selected model levels in W1-1000 than W1-Ope. Different from the consistently degraded global wind forecasts, W1-1000 shows slightly better temperature forecasts than W1-Ope at several lower levels within 2 days and at 10 hPa beyond 2 days. This may suggest that the optimal localization length scale varies with model variable. Compared to W1-Ope, W1-300 degrades global temperature forecasts above 150 hPa. This is hypothesized to result from the negative impacts of applying too-tight horizontal localization at such levels in W1-300 relative to W1-Ope. Compared to the global wind forecasts, the more degraded global temperature forecasts above 150 hPa in W1-300 may also suggest that the optimal horizontal localization length scale is variable-dependent. Below 150 hPa, W1-300 produces slightly more accurate or comparable temperature forecasts relative to W1-Ope. However, W1-300 shows more accurate global wind forecasts up to 50 hPa at shorter forecast lead times. This is consistent with the 6-h background forecast verification (see Fig. 7). This again suggests the potential additional benefits of slightly reducing horizontal localization length scale at lower model levels in the operational hybrid 4DEnVar system for the global wind forecasts. Different from W1-1000 and W1-300, W2-NoCross and W2-Cross that apply SDL consistently improve both global temperature and wind forecasts over W1-Ope at most model levels and forecast lead times. This result suggests SDL may implicitly produce overall horizontal localization that is level- and variable-dependent. This effect may be attributed to that the relative variance projected on different scale wave bands varies with variables and model levels. It is noticed that W2-NoCross shows slightly degraded temperature forecasts relative to W1-Ope below 850 hPa between 3 and 4 days. This may suggest that the current scale separation based on the wind variable may not be optimal for the temperature variable, especially when applying SDL-NoCross.

Figures 9a and 9b further compare the 5-day global forecasts between W2-NoCross and W2-Cross. W2-NoCross tends to show slightly better global forecasts than W2-Cross within 1-day forecast lead times. Beyond 1 day, however, W2-Cross improves the global forecasts over W2-NoCross. The advantage of

W2-NoCross at shorter forecast lead times may be associated with the effective local spatial averaging of ensemble covariances in SDL-NoCross, due to its neglect of cross-wave-band covariances. While the local spatial averaging in SDL-NoCross may help alleviate sampling error and improve fitting of the analysis to the observations, it would possibly retain less heterogeneity of ensemble covariances (Caron et al. 2019). In contrast, SDL-Cross partially includes the cross-wave-band covariances and retains more heterogeneity of ensemble covariances. As detailed later in section 6b, W2-Cross analysis is more balanced than W2-NoCross. The retained heterogeneity of ensemble covariances in W2-Cross, together with its more balanced analysis may explain its more accurate global forecasts at longer forecast times than W2-NoCross.

e. TC track forecast verification

As discussed in the introduction, tropical cyclone (TC) track prediction requires that the TC, its embedded large-scale environment and their interaction to be analyzed properly. Therefore, we further investigate how SDL influences the TC track forecasts.

In the cycled DA experimental period, there were 15 named TCs in the North Atlantic basin and North Pacific basin. Ten TCs reached typhoon or hurricane category (Fig. 10). The NCEP TC tracker (Marchok 2002) was used to calculate the forecasted TC location. Following Wang and Lei (2014) and Huang and Wang (2018), the TC track error was calculated against the best track data[2] and averaged over all 15 TCs. The statistical significance tests were performed for paired track error differences among all the experiments. In the TC track forecast verification (Figs. 11 and 13), we selected W2-Cross as the referencing experiment. This selection aims to facilitate displaying statistical significance results for the following three comparisons: SDL versus scale-invariant localization, SDL-NoCross versus SDL-Cross, and two- versus three-wave-band SDL experiments. Figure 11a shows the TC track error difference relative to W2-Cross as a function of forecast lead time. W1-Ope improves the TC track forecasts over W1–1000 almost out to 5 days, and over W1-300 beyond 3 days. On the other hand, W2-Cross and W2-NoCross show significantly improved TC track forecasts over W1-1000 out to 4 days, and over W1-300 beyond 3 days. These results suggest improved analyses in W2-Cross and W2-NoCross relative to W1-1000 and W1-300 lead to improved subsequent TC track forecasts. Furthermore, W2-Cross improves the TC track forecasts over W2-NoCross beyond 3 days. This is consistent with better global forecasts in W2-Cross relative to W2-NoCross at longer forecast lead times in Figs. 9a and 9b. However, the TC track forecasts between W1-Ope and W2-Cross are statistically indistinguishable. This is not the case in their global forecast comparison in Figs. 8g and 8h. It is hypothesized that the current scale separation in the two-wave-band SDL experiments may not be sufficient to improve the TC track forecasts. The impact of SDL on TC

---

[2] The best TC track data are available at https://www.nhc.noaa.gov/data/#hurdat and https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks.
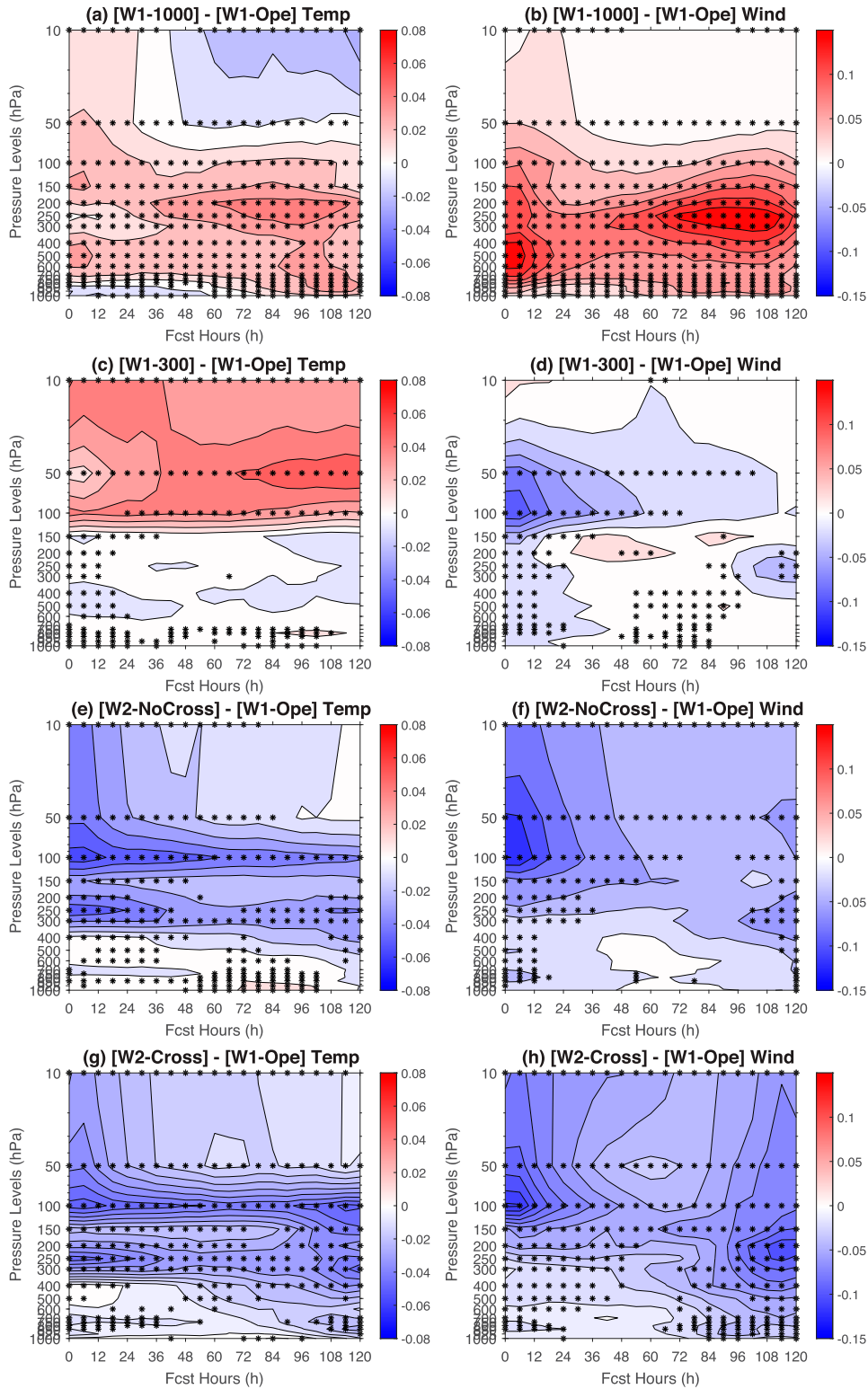
FIG. 8. Globally and temporally averaged root-mean-square-error (RMSE) difference from W1-Ope for the (left) temperature (unit: K) and (right) wind (unit: m s^{-1}) forecasts in (a),(b) W1-1000; (c),(d) W1-300; (e),(f) W2-NoCross; and (g),(h) W2-Cross that were verified against ERA-Interim as a function of forecast time out to 5 days on the horizontal axis and pressure level on the vertical axis. Blue (red) color indicates the improved (degraded) forecasts relative to W1-Ope. The asterisks at the corresponding forecast times and pressure levels indicate that the RMSE difference is statistically significant by applying the paired permutation test with the false discovery method at 5% significance level.
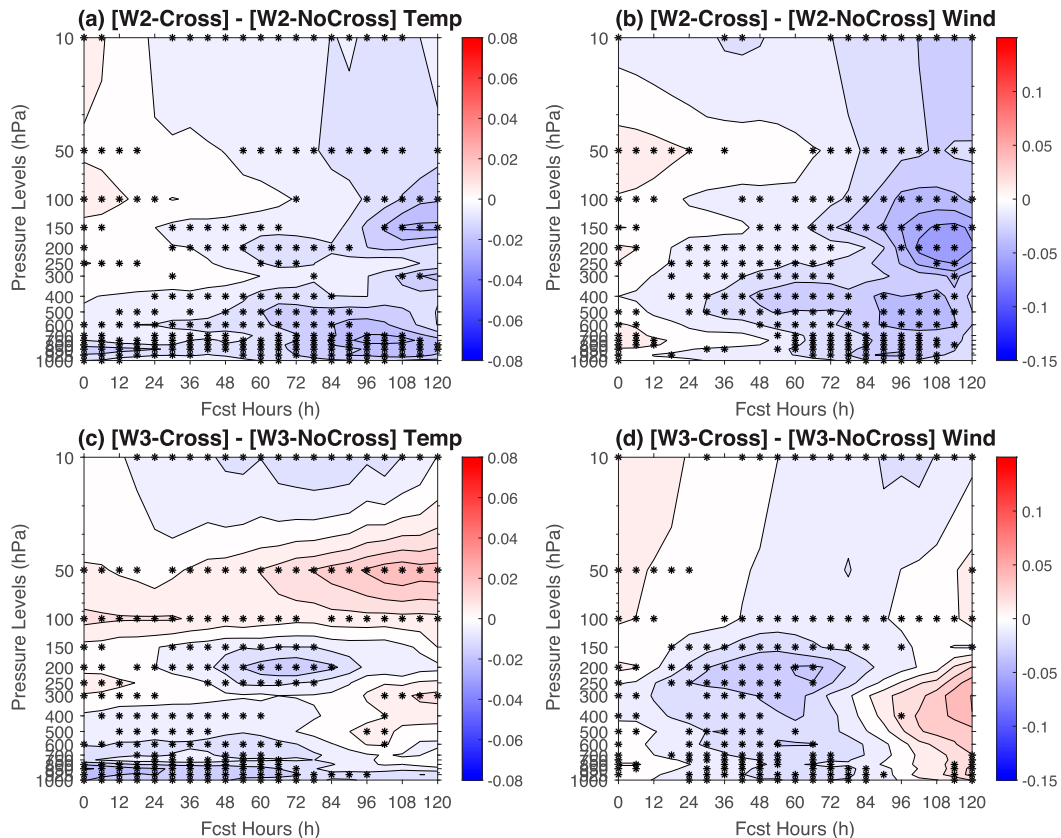
FIG. 9. As in Fig. 8, but for the global (left) temperature and (right) wind forecast RMSE difference between the experiments applying SDL-Cross and SDL-NoCross in the (a),(b) two- and (c),(d) three-wave-band SDL experiments. Blue (red) color indicates SDL-Cross shows improved (degraded) forecasts compared to SDL-NoCross.

track forecasts will be further examined in the three-wave-band SDL experiments in section 5b. The percentages of the track forecasts that are more accurate than W2-Cross (Zapotocny et al. 2008; Wang and Lei 2014) in Fig. 11b are consistent with the TC track error differences in Fig. 11a. For instance, at the forecast lead times beyond 2.5 days, more than 50% of the track forecasts in W2-Cross are more accurate than W1-1000, W1-300, and W2-NoCross.

5. Comparison of two- and three-wave-band SDL experiments

To investigate how SDL performs in response to the number of decomposed wave bands, the two- and three-wave-band SDL experiments are further compared in this section. As discussed in the introduction, the main difference in the two- and three-wave-band SDL experiment designs is tighter horizontal localization applied at medium-scale wave band in the three-wave-band SDL experiments. Consistent with the two-wave-band SDL experiments, the three-wave-band SDL experiments improve the global temperature and wind forecasts over W1-Ope at most model levels and forecast lead times (not shown here). This section will focus on the comparison among the two- and three-wave-band SDL experiments.

a. Global forecast verification against ERA-Interim

Figure 12 shows the RMSE difference of the global temperature and wind forecasts against ERA-Interim between the two- and three-wave-band SDL experiments. Above 50 hPa, the three-wave-band SDL experiments produce generally less accurate global forecasts over 5 days than their two-wave-band SDL experiment counterparts. As discussed earlier, a wider overall horizontal localization length scale is beneficial at upper model levels. Reduced overall horizontal localization length scales in the three-wave-band SDL experiments, due to tighter horizontal localization applied at medium-scale wave band, may be the reason for the degraded global forecasts at upper model levels. On the other hand, a tighter horizontal localization length scale is beneficial overall at lower levels. So an overall reduction in the horizontal localization length scale in the three-wave-band SDL experiments may explain the better global temperature and wind forecasts below 100 hPa, at least out to 3 days compared to the two-wave-band SDL experiments. In particular, the advantage of W3-NoCrsoss relative to W2-NoCross below 100 hPa lasts for nearly the entire 5 days, while that of W3-Cross relative to W2-Cross only lasts out to 3 days. As detailed later in section 6b, the analysis in the three-wave-band SDL experiments is more imbalanced than the two-wave-band SDL experiment counterparts. Furthermore,

(a) Best track in ATL

(b) Best track in WPC
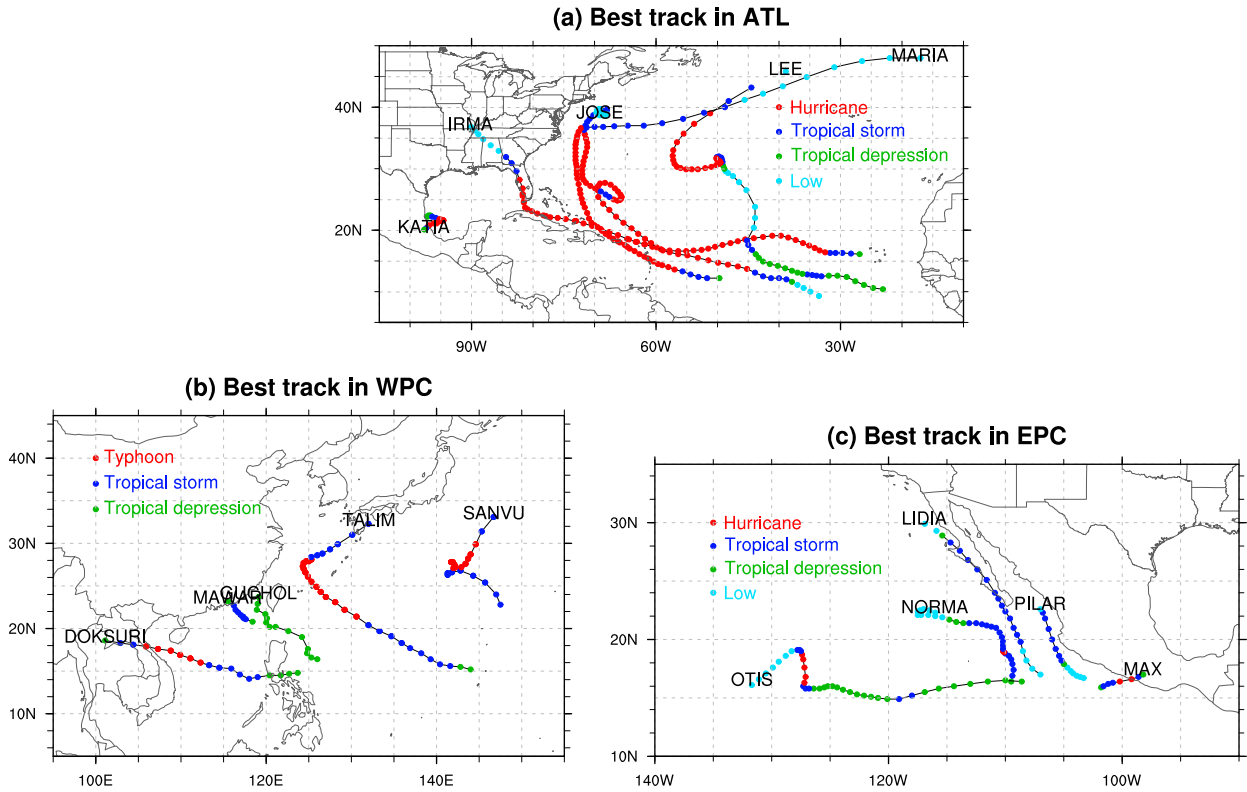

(c) Best track in EPC

FIG. 10. Best track of the tropical cyclones during the experiment period in the (a) Atlantic, (b) western Pacific, and (c) eastern Pacific basins.

the increased imbalance from W2-Cross to W3-Cross is larger than that from W2-NoCross to W3-NoCross. So it is likely that the negative impacts from less balanced analysis in W3-Cross relative to W2-Cross overwhelm the positive benefits of more accurate analysis resultant from its overall tighter horizontal localization. This may explain why the improved forecasts of W3-Cross versus W2-Cross below 100 hPa only last up to 3 days.

(a) Track err. diff. from W2-Cross

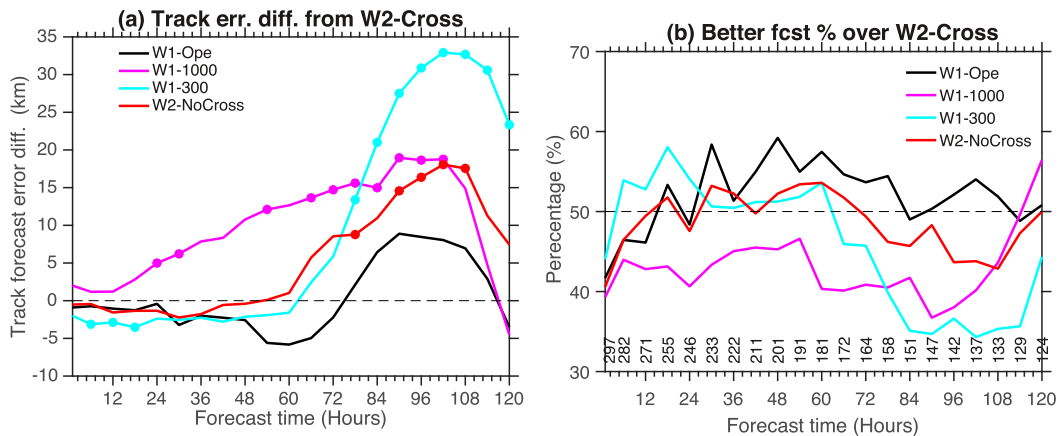## (b) Better fcst % over W2-Cross

FIG. 11. (a) Track forecast error difference in W1-Ope (black), W1-1000 (magenta), W1-300 (cyan), and W2-NoCross (red) relative to W2-Cross. The bold dots on the curves in (a) indicate that the track error difference is statistically significant by applying the paired permutation test with the false discovery method at 5% significance level at the corresponding forecast lead time. (b) Percentage of the track forecasts that are more accurate than W2-Cross with the same line colors in (a). The numbers right above the x axis in (b) denote the sample size at the corresponding forecast lead time.
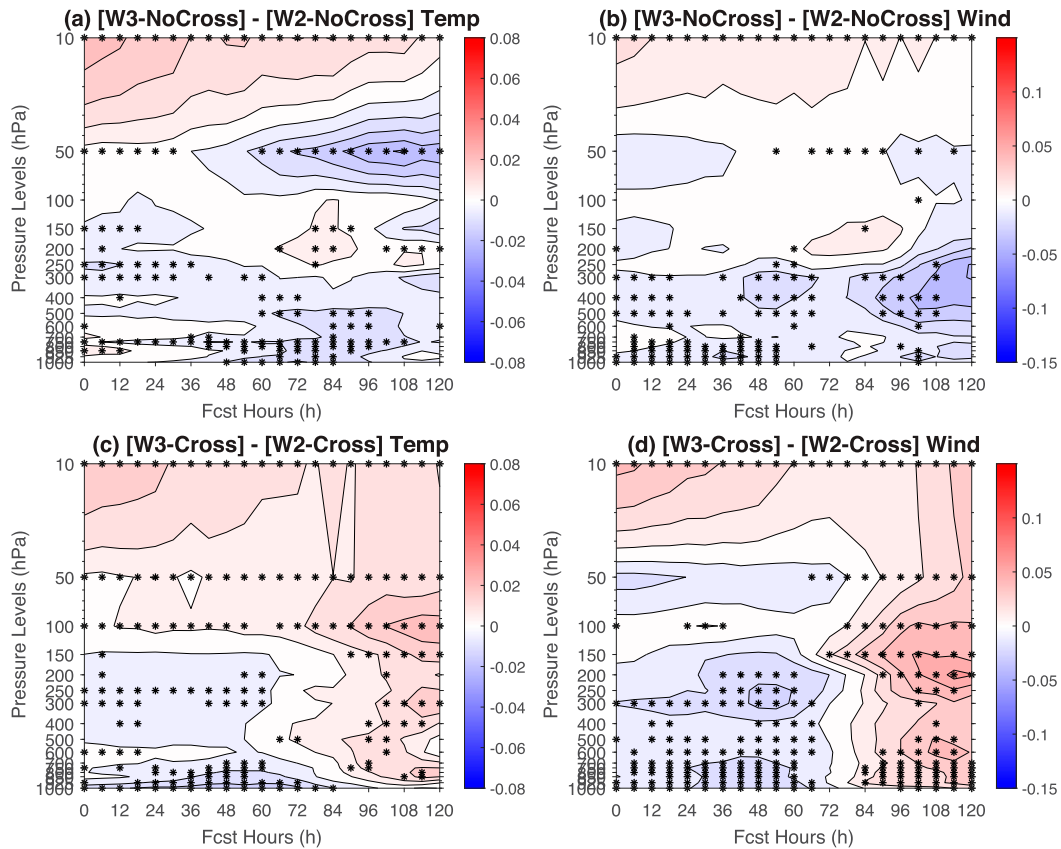
FIG. 12. As in Fig. 8, but for the global (left) temperature and (right) wind forecast RMSE difference between the two- and three-wave-band SDL experiment counterparts that apply (a),(b) SDL-NoCross and (c),(d) SDL-Cross. Blue (red) color indicates the three-wave-band SDL experiment shows improved (degraded) forecasts compared to its two-wave-band experiment counterpart.

Figures 9c and 9d show the forecast RMSE difference between W3-NoCross and W3-Cross. As in the two-wave-band SDL experiments, W3-NoCross produces slightly more accurate global wind forecasts within 12 h than W3-Cross. As the forecast lead time increases, the global wind forecasts in W3-Cross become more accurate than W3-NoCross. This variation of the wind forecast performance for W3-NoCross relative to W3-Cross as the forecast lead time increases also happens to the global temperature forecasts between 400 and 150 hPa, and at about 10 hPa. As discussed in the two-wave-band SDL experiments in section 4d, the better global forecasts of W3-Cross over W3-NoCross at longer forecast lead times may be associated with retained heterogeneity of ensemble covariances and more balanced analysis. However, it is also noted that the outperformance of W3-Cross over W3-NoCross lasts for a shorter period of time than that of W2-Cross over W2-NoCross. This may be related to less amount of imbalance reduction between SDL-Cross and SDL-NoCross when increasing from two wave bands to three wave bands. Between 100 and 50 hPa, W3-Cross shows worse global temperature forecasts than W3-NoCross over 5 days. However, this is neither the case between W2-NoCross and W2-Cross, nor the case for the wind forecasts. These results suggest that the relative

performance between SDL-NoCross and SDL-Cross, owing to the partial inclusion of cross-wave-band covariances in SDL-Cross, could vary with the number of decomposed wave bands in SDL, model level, and model variable.

b. TC track forecast verification

Figure 13 shows the TC track forecast error difference in the two- and three-wave-band SDL experiments relative to W2-Cross, and the percentages of the TC track forecasts that are better than W2-Cross. W1-Ope is included in Fig. 13 for further comparison with the three-wave-band SDL experiments for the TC track forecasts. W3-Cross significantly improves the TC track forecasts over W1-Ope and W2-Cross within 2-day forecast lead times. W3-NoCross in general shows more accurate TC track forecasts than W2-NoCross between 1.5 and 3 days. Improved TC track forecasts in the three-wave-band SDL experiments over their two-wave-band SDL counterparts may benefit from the three-wave-band scale separation and tighter horizontal localization length scale applied at medium-scale wave band. For example, the decomposed wind background displays more representative and distinguished features in the three-wave-band SDL experiments (Fig. 3). However, W3-Cross shows statistically less accurate TC track
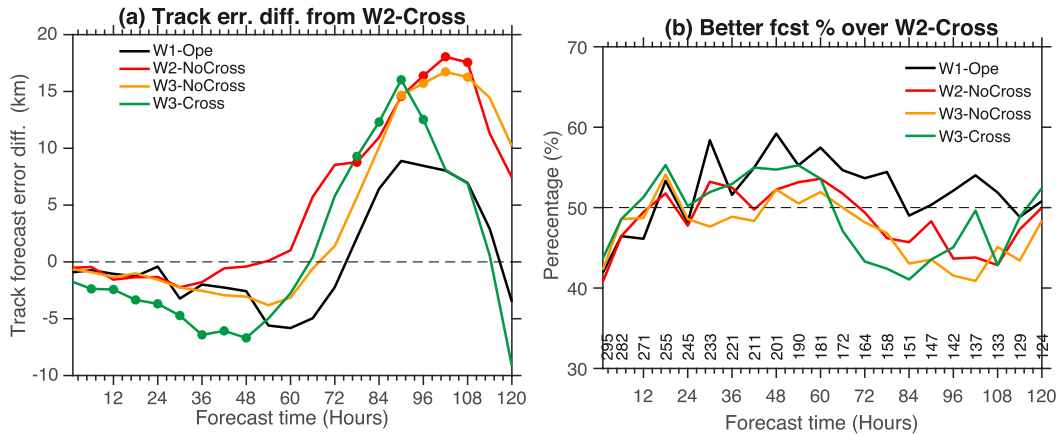
FIG. 13. As in Fig. 11, but for (a) the track forecast error difference and (b) percentage of more accurate track forecasts in W1-Ope (black), W2-NoCross (red), W3-NoCross (orange), and W3-Cross (green) in contrast to W2-Cross.

forecast than W2-Cross between 3 and 4 days. This is consistent with its generally less accurate global forecasts in W3-Cross than W2-Cross at such forecast time periods. Furthermore, the percentage differences in Fig. 13b are generally consistent with the TC track error differences in Fig. 13a. For instance, more than 50% of track forecasts in W3-Cross are more accurate than W2-Cross between 12 h and 2.5 days.

6. Additional diagnostics to understand performance differences

a. Localized correlation pattern comparison

To have a more systematic evaluation of localized correlation patterns from all of the localization methods, a total of 200 localized meridional wind correlation samples at 500 hPa were collected over the globe for all eight experiments. Figure 14 shows 45 localized correlation samples. In general, the localized correlation patterns vary with latitude, geographical location and weather system (Buehner 2012). As expected, a tighter horizontal localization length scale produces more localized and smooth ensemble correlation patterns in the scale-invariant localization experiments. W1-1000 applying the most wide horizontal localization length scale shows the broadest localized correlation patterns with the most heterogeneity and largest amount of noise. In contrast, W1-650 shows narrower and less noisy localized correlation patterns than W1-1000. W1-300 shows the tightest and most smooth localized correlation patterns. Slightly broader localized correlation patterns in W1-Ope relative to W1-300 is because W1-Ope applies a slightly wider horizontal localization length scale at 500 hPa (Fig. 2). Due to a smaller horizontal localization length scale applied at small- and/or medium-scale wave bands, the localized correlation patterns in the SDL experiments are tighter, smoother and retain less heterogeneity of ensemble correlations compared to W1-1000. On the other hand, the SDL experiments show broader localized correlation patterns than W1–300 due to their wider horizontal localization

length scales applied at large- and/or medium-scale wave bands. Owing to the local spatial averaging of ensemble covariances, the localized correlation patterns from SDL-NoCross tend to be less heterogenetic than those from SDL-Cross. By applying a tighter horizontal localization length scale at medium-scale wave band, the three-wave-band SDL experiments produce more localized correlation patterns than their two-wave-band SDL counterparts.

To further quantify and compare the spatial structure and variation of localized ensemble correlations, the mean and standard deviation of 200 correlation samples at 500 hPa were calculated and shown as a function of distance in Fig. 15. Consistent with Fig. 14, a tighter horizontal localization length scale shows a narrower averaged localized correlation structure in the scale-invariant localization experiments. By applying the horizontal localization length scale of 300 km e-folding distance at small-scale wave band, the two- and three-wave-band SDL experiments show similar averaged localized correlations as W1-300 at distances within 350 km. However, at distances between 600 and 1500 km, the two- and three-wave-band SDL experiments even show larger averaged localized correlations than W1-1000. An inspection of nonlocalized ensemble correlations at different wave bands shows that the decomposed nonlocalized small-scale correlations have negative values at such intermediate distances (not shown). Compared to W1-1000 that applies a much wider horizontal localization length scale at all scales, the tighter horizontal localization length scales applied at small- and/or medium-scale wave bands in the two- and three-wave-band SDL experiments would constrain more or even completely eliminate these nonlocalized small-scale negative correlations, thus resulting in their overall larger localized ensemble correlations at such intermediate distances. Therefore, SDL could produce an overall effective localization function having a different shape from scale-invariant localization (e.g., non-Gaussian structure in SDL versus Gaussian structure in scale-invariant localization). The three-wave-band SDL experiments, applying a tighter horizontal localization length scale at medium-scale
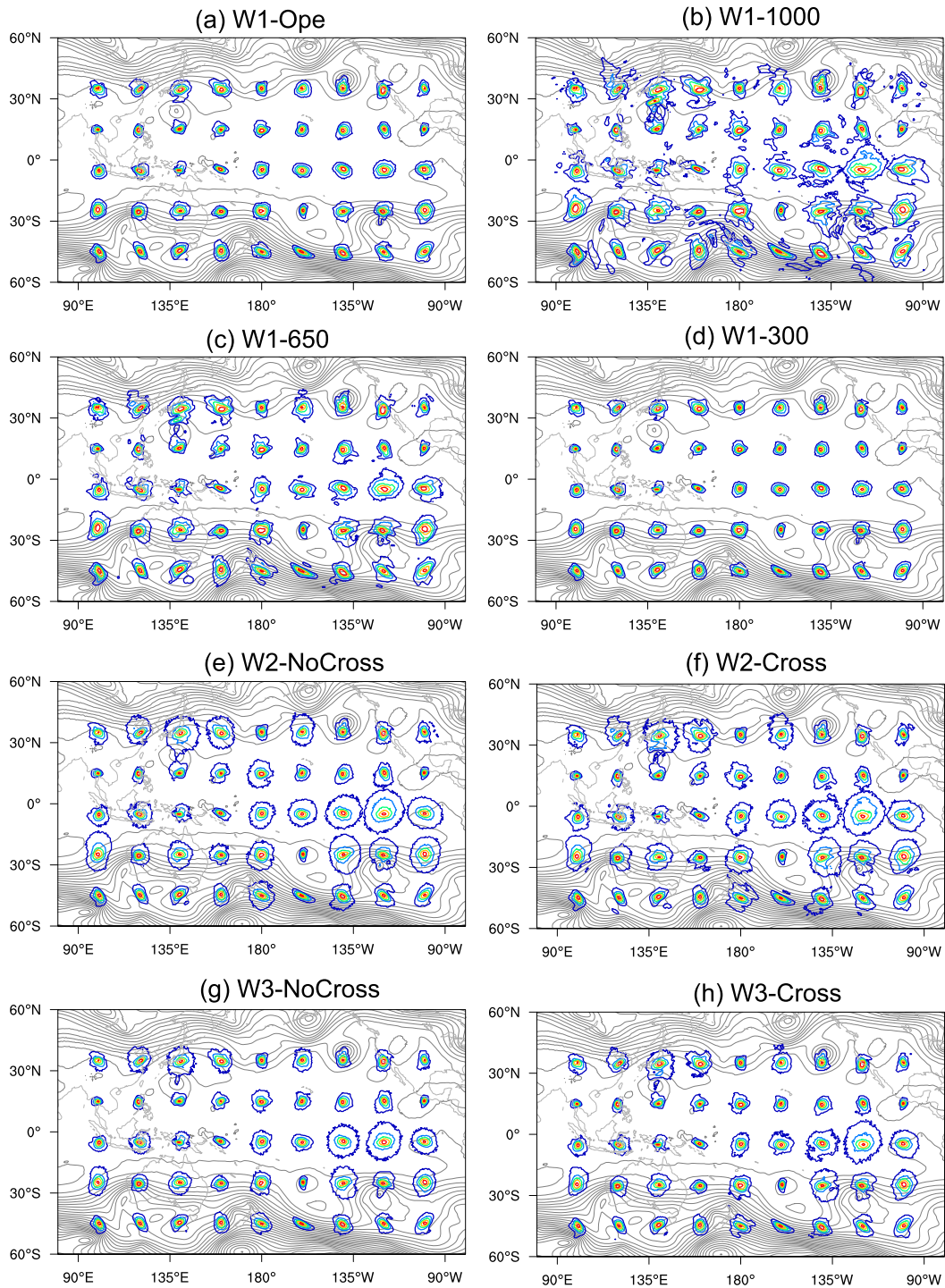
FIG. 14. 45 localized meridional wind correlation samples at 500 hPa distributed over the globe in (a) W1-Ope, (b) W1-1000, (c) W1-650, (d) W1-300, (e) W2-NoCross, (f) W2-Cross, (g) W3-NoCross, and (h) W3-Cross. The thin gray contours give the geopotential height at 500 hPa. The colored contours display the correlation magnitude from 0.15 to 0.95 with an interval of 0.2. The gray contours denote the geopotential height with an interval of 40 gpm at 500 hPa.
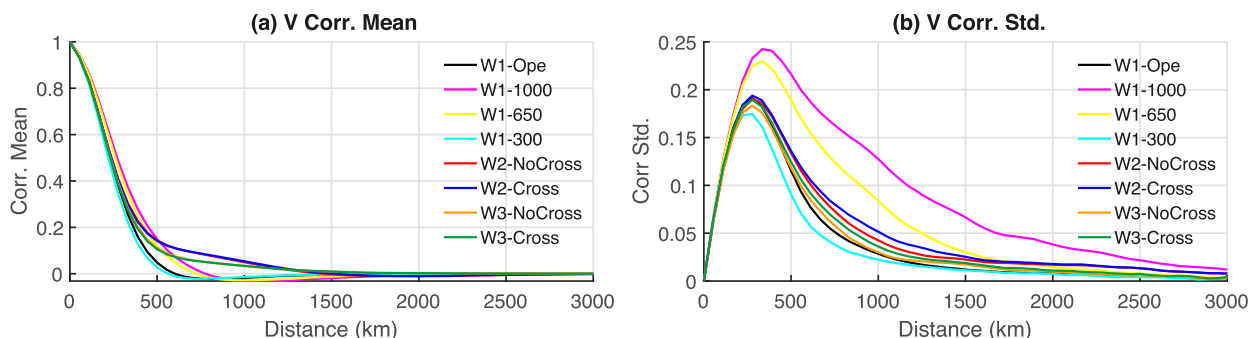
FIG. 15. (a) Mean and (b) standard deviation of 200 samples of localized meridional wind correlations as a function of distance on the horizontal axis that were collected from W1-Ope (black), W1-1000 (magenta), W1-650 (yellow), W1-300 (cyan), W2-NoCross (red), W2-Cross (blue), W3-NoCross (orange), and W3-Cross (green) in Fig. 14.

wave band, show smaller averaged localized correlations at distances between 350 and 1200 km. Due to the averaging of 200 correlation samples, the averaged localized correlation differences in SDL-Cross and SDL-NoCross are almost indistinguishable, which is not true in their case-by-case comparison in Fig. 14.

As for the standard deviation of the localized correlations, W1-1000 and W1-650 shows much larger standard deviation than W1-300 and W1-Ope. This suggests that tighter localization retains less spatial variation of raw ensemble correlations. By applying tighter horizontal localization at small- and/or medium-scale wave bands, all of the SDL experiments have smaller standard deviation than W1-1000. Due to applying a broader horizontal localization length scale at large- and/or medium-scale wave bands, the SDL experiments show larger standard deviation than W1-300 beyond 300 km. However, within 300 km, W2-Cross and W3-Cross show comparable standard deviation as W1-300, while W2-NoCross and W3-NoCross have smaller standard deviation than W1-300. More noticeably, the experiments applying SDL-Cross that partially includes cross-wave-band covariances consistently show larger standard deviation than those applying SDL-NoCross. This may be an evidence that SDL-Cross partially including cross-wave-band covariances contains higher degrees of heterogeneity than SDL-Cross as discussed in the introduction. Finally, the three-wave-band SDL experiments show less standard deviation or variation than their two-wave-band SDL experiment counterparts at distances between 300 and 1700 km due to the effective tighter localization in three band. Consistent with the single observation experiments, W2-NoCross and W2-Cross show large differences compared to W1-650 with respect to localized correlation pattern and spatial variation in Figs. 14 and 15. This further suggests that SDL is not equivalent to applying scale-invariant localization with an intermediate localization length scale.

b. Impact of dynamical balance

Covariance localization causes imbalance in the resultant analyses in the ensemble-based DA system (Wang et al. 2007a; Holland and Wang 2013; Wang et al. 2013; Lei and Whitaker 2016; Caron et al. 2019). To measure the imbalance, the

absolute hourly pressure tendency (Lynch and Huang 1992) was calculated for all the seven experiments (Fig. 16). Statistical significance results from the paired t-test applied in Fig. 16 show that imbalance difference in the subsequent comparison of paired experiments is statistically significant at 95% confidence level. The operationally tuned, level-dependent horizontal localization length scale in W1-Ope is expected to better reflect and maintain the dominating scale at each model level compared to W1-1000 and W1-300 that use a single fixed horizontal localization length scale at all model levels. This may lead to the least imbalance in W1-Ope among all the scale-invariant localization experiments. W1-300 is less balanced than W1-Ope, consistent with applying tighter horizontal localization length scales and the imbalances that may result. However, W1-1000 is less balanced than W1-Ope and W1-300. This seemingly out-of-expectation result is consistent with Greybush et al. (2011) who hypothesized that larger analysis increments that are required to correct an inaccurate background have the potential of producing more imbalance through the DA update. As shown in Figs. 6 and 7, W1-1000 generally shows the worst background forecasts, but the largest analysis increments compared to W1-Ope and W1-300.

The analyses in the SDL experiments are less balanced than W1-Ope, which could be because the overall effective horizontal localization length scale in the SDL experiments is smaller especially at upper model levels. Interestingly, the analysis in SDL-Cross is more balanced than SDL-NoCross, likely due to the partial maintenance of the cross-wave-band covariances in SDL-Cross. The more balanced analysis in SDL-Cross may contribute to its generally better forecasts at longer forecast lead times than SDL-NoCross in section 5. Furthermore, the three-wave-band SDL experiments are less balanced than their two-wave-band SDL experiment counterparts. This may be the result of reduced overall effective horizontal localization length scales in the three-wave-band SDL experiments.

c. Forecast error comparison as a function of total wavenumber

To investigate how SDL affects forecast error at different scales, the errors of the global forecasts against ERA-Interim
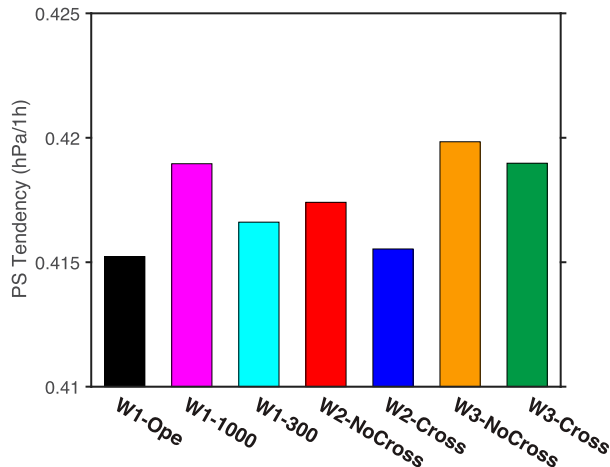
Fig. 16. Globally averaged absolute hourly surface pressure tendency (unit: hPa 1 h^{-1}) in W1-Ope (black), W1-1000 (magenta), W1-300 (cyan), W2-NoCross (red), W2-Cross (blue), W3-NoCross (orange), and W3-Cross (green). The absolute hourly surface pressure tendency was calculated after the four-dimensional incremental analysis update (4DIAU) was complete, and averaged over the last 4 weeks of the full cycled period.

in Fig. 8 were decomposed in spectral space and displayed as a function of total wavenumber. A representative total energy norm was performed on decomposed errors (Wang and Bishop 2003), and averaged over all selected levels in Fig. 8. Figure 17 shows the total energy error difference relative to W1-Ope as a function of forecast lead time and total wavenumber. By applying a wider horizontal localization length scale below 100 hPa, W1-1000 (Fig. 17a) shows larger error total energy error than W1-Ope at most total wavenumbers and forecast lead times. The smaller total energy error of W1-1000 versus W1-Ope at total wavenumbers below 10 within 2 days may be because a wider horizontal localization length scale contributes positively to the large-scale component of ensemble covariances and resultant analysis. In comparison, since tighter localization is beneficial for the small-scale component of ensemble covariances and resultant analysis, W1-300 (Fig. 17b) generally shows smaller error than W1-Ope at total wavenumbers above 10. At total wavenumbers below 10, W1-300 shows comparable or slightly larger error total energy than W1-Ope.

In addition, the two- and three-wave-band SDL experiments show smaller total energy error than W1-Ope at most total wavenumbers and forecast lead times, especially for the experiments applying SDL-Cross. Their largest total energy error reduction relative to W1-Ope appears at total wavenumbers between 5 and 20 beyond 3 days where W1-Ope shows the maximum error total energy (not shown here). Without considering cross-wave-band covariances, the experiments applying SDL-NoCross tend to have slightly larger total energy error than those applying SDL-Cross at longer forecast lead times. For example, the experiments applying SDL-NoCross show slightly larger total energy error than W1-Ope at total wavenumbers of 3 and 4 beyond 2 days. However, this is not the case

for the experiments applying SDL-Cross, suggesting additional benefits of partially including cross-wave-band covariances. By applying a tighter horizontal localization for the medium-scale wave band, the three-wave-band SDL experiments generally show slightly larger total energy error than their two-wave-band SDL counterparts, especially at total wavenumbers between 5 and 20 that correspond to the error total energy maximum in W1-Ope. Overall, the SDL experiments, especially when including cross-wave-band correlations (SDL-Cross), show the potential to improve global forecasts over W1-Ope for nearly all total wavenumbers in contrast to a subset of total wavenumbers in W1-1000 and W1-300.

7. Computational cost comparison

Table 2 summarizes the computational cost in wall clock time in each of the four components in a single 4DEnVar DA cycle. Further extending control variables in SDL increases the amount of computation in the 4DEnVar update and total computational cost, especially when a larger number of wave bands are adopted in SDL. In general, the ensemble forecasts are the most expensive component in an EnVar DA cycle. Therefore, the two-wave-band and three-wave-band SDL experiments only increase the total computational cost by 14% and 28%, respectively. Given that the two- and three-wave-band SDL experiments significantly improve the global forecasts to 5 days over the stricter reference W1-Ope and that W3-Cross even shows more accurate TC track forecasts than W1-Ope at shorter forecast lead times, SDL shows promise to be implemented operationally.

8. Conclusions and discussion

Two SDL variants, with (SDL-Cross) and without (SDL-NoCross) considering cross-wave-band covariances were formulated based on the full **B**-preconditioned EnVar, and implemented in the GSI-based 4DEnVar system by further extending control variables. SDL performs a single-step simultaneous assimilation of all available observations, while applying different amounts of localization to different scales of ensemble covariances. The complete removal of cross-wave-band covariances in SDL-NoCross results in a local spatial averaging of ensemble covariances (Buehner and Charron 2007) and retains less heterogeneity of ensemble covariances (Caron et al. 2019). SDL-Cross partially includes cross-wave-band covariances and retains more heterogeneity of ensemble covariances than SDL-NoCross. The performances of SDL-NoCross and SDL-Cross were evaluated for general global forecasts and TC track forecasts in the FV3-based GFS through 1-month cycled DA experiments.

The two-wave-band SDL experiments improve global forecasts to nearly 5 days over W1-1000 and W1-300 applying scale-invariant, level-invariant localization, and even over W1-Ope, which applies operationally tuned, scale-invariant, level-dependent localization. By applying a much wider horizontal localization length scale, W1-1000 generally degrades global forecasts below 50 hPa relative to W1-Ope. By applying a tighter horizontal localization scale, W1-300 shows worse
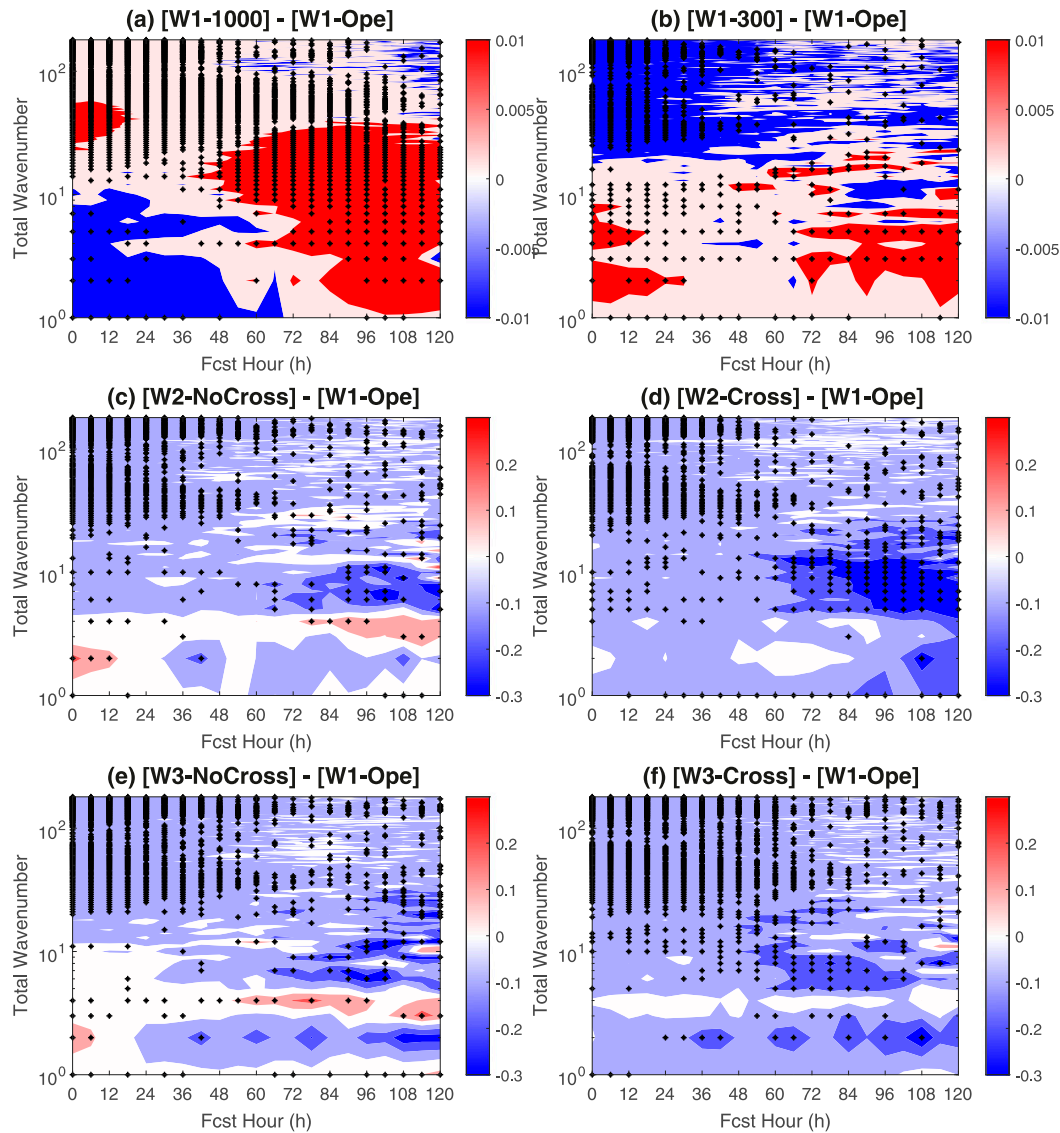
FIG. 17. Power spectra of error total energy difference relative to W1-Ope in (a) W1-1000, (b) W1-300, (c) W2-NoCross, (d) W2-Cross, (e) W3-NoCross, and (f) W3-Cross, as a function of forecast time to 5 days on the horizontal axis and total wavenumber on the vertical axis. See the text for details of error total energy calculation. Blue (red) color indicates smaller (larger) error total energy relative to W1-Ope. The asterisks at the corresponding forecast times and total wavenumbers indicate that the error total energy difference is statistically significant by applying the paired permutation test with the false discovery method at 5% significance level. Note that different color bars are applied in (a),(b) vs (c)–(f) to more clearly show statistically significant positive error total energy differences in (a) and (b).

global forecasts than W1-Ope at upper levels, especially for temperature forecasts. As for TC track forecasts, the two-wave-band SDL experiments outperform W1-1000 out to 4 days, and over W1-300 beyond 3 days. W1-Ope generally produces more accurate TC track forecasts than W1-1000 and W1-300.

To examine how SDL performs with respect to the number of decomposed wave bands, the three-wave-band SDL experiments were compared with the two-wave-band SDL experiments. In their design, the three-wave-band SDL experiments apply tighter horizontal localization at medium-scale wave

band compared to the two-wave-band SDL experiments. This leads to overall tighter effective horizontal localization compared to the two-wave-band SDL experiments. As suggested in the cycled experiment results and the operationally tuned, level-dependent, scale-invariant horizontal localization settings, a wider horizontal localization length scale is beneficial at upper levels, while a tighter horizontal localization makes positive contribution at lower levels. Therefore, the reduced effective horizontal localization length scale in the three-wave-band SDL experiments may explain the generally degraded

TABLE 2. Wall clock time in minutes for each component in a single 4DEnVar DA cycle. The same number of 1260 cores were run in each component for all experiments.

	Wall clock time (min) in each component in a single 4DEnVar DA cycle					
Expt	EnVar update	EnKF update	Control background forecasts	Ensemble background forecasts	Total wall clock time (min)	Total cost ratio relative to W1
W1	15	7	3	45	70	1.0
W2	25				80	1.14
W3	35				90	1.28

global forecasts above 50 hPa and better global forecasts below 100 hPa compared to the two-wave-band SDL counterparts. On the other hand, the degraded global forecasts of W3-Cross versus W2-Cross below 50 hPa at longer forecast lead times may be the result of the advantage of reduced overall effective horizontal localization length scale being overwhelmed by more imbalance in the analysis. Compared to statistically indistinguishable TC track forecasts between W2-Cross and W1-Ope, W3-Cross shows significantly improved TC track forecasts than W1-Ope and W2-Cross within 2 days. This suggests the important role of scale separation in the SDL implementation, especially for TC track forecasts. This should be explored in the future to further improve TC track forecasts in the SDL experiments relative to W1-Ope.

Due to the local spatial averaging of ensemble covariances, SDL-NoCross tends to show slightly better global forecasts than SDL-Cross at shorter forecast lead times. At longer forecast lead times, SDL-Cross outperforms SDL-NoCross for the global forecasts, especially in the two-wave-band SDL experiments. However, comparable performance between SDL-NoCross and SDL-Cross were found in Caron et al. (2019) that run a 75-member ensemble populated by time-lagged method in a regional 3DEnVar system. The better forecast performance of SDL-Cross than SDL-NoCross at longer forecast lead times in this study may benefit from its more accurate estimate of cross-wave-band covariances by directly running and updating an 80-member ensemble, higher degrees of heterogeneity of ensemble covariances and more balanced analysis in SDL-Cross. The TC track forecasts between SDL-Cross and SDL-NoCross are generally consistent with their global wind forecasts at lower model levels. For example, W2-Cross shows significantly better TC track forecasts than W2-NoCross at longer forecast lead times.

Due to further extending control variables, the two- and three-wave-band SDL experiments increase total computational cost by 14% and 28%, respectively, compared to the scale-invariant localization experiments. However, SDL shows statistically significantly improved global forecasts and the potential for improving TC track forecasts over scale-invariant localization. Moreover, SDL without requiring additional ensemble forecasts is computationally much cheaper than directly increasing ensemble size. Therefore, SDL shows the potential to be implemented operationally.

By comparing localized ensemble correlations of W1-650 and two-wave-band SDL experiments, it suggests that SDL is not equivalent to applying scale-invariant localization with an intermediate localization length scale. In our current SDL experiments, the localization length scale is level invariant. However, the two-wave-band SDL experiments showed increasing effective ensemble correlation length scales at large- and small-scale wave bands as model level increases (not shown here). Two-wave-band level-dependent SDL experiments were motivated by increasing the horizontal localization length scales beyond 1000 and 300 km e-folding distances for the large- and small-scale wave band, respectively, above 0.3 sigma model level following the increasing trend in W1-Ope (not shown here). However, compared to the two-wave-band level-invariant SDL experiments, these two-wave-band level-dependent SDL experiments showed degraded global forecasts above 200 hPa, but comparable global forecasts below 200 hPa and comparable TC track forecasts. The horizontal localization length scales in the two-wave-band level-dependent SDL experiments may need more tuning. Following Caron and Buehner (2018), the objective method proposed by Ménétrier et al. (2015) was attempted to determine optimal horizontal localization length scale at each wave band in our SDL experiments, but it showed worse global forecasts than our current SDL experiments. This may be related to the assumption of independent members in this objective method, which may be not true in the operational NWP applications (Caron and Buehner 2018). Research on seeking the optimal effective localization length scale in SDL would be further explored in the future. In addition to applying SDL on the horizontal direction, further development of SDL on the vertical direction is ongoing in the GSI-based global 4DEnVar system. Furthermore, encouraging SDL results presented here also warrant its additional tests in the operational global hybrid 4DEnVar where a higher model resolution is used.

Data availability statement. All model data produced during this study have been archived locally and are available upon request to the corresponding author.

REFERENCES

Anderson, J., and L. Lei, 2013: Empirical localization of observation impact in ensemble Kalman filters. *Mon. Wea. Rev.*, **141**, 4140–4153, https://doi.org/10.1175/MWR-D-12-00330.1.

Bishop, C. H., and D. Hodyss, 2009: Ensemble covariances adaptively localized with ECO-RAP. Part 2: A strategy for the atmosphere. *Tellus*, **61A**, 97–111, https://doi.org/10.1111/j.1600-0870.2008.00372.x.

——, B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bloom, S. C., L. L. Takacs, A. M. Da Silva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256–1271, https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2.

Buehner, M., 2005: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quart. J. Roy. Meteor. Soc.*, **131**, 1013–1043, https://doi.org/10.1256/qj.04.15.

——, 2012: Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Mon. Wea. Rev.*, **140**, 617–636, https://doi.org/10.1175/MWR-D-10-05052.1.

——, and M. Charron, 2007: Spectral and spatial localization of background-error correlations for data assimilation. *Quart. J. Roy. Meteor. Soc.*, **133**, 615–630, https://doi.org/10.1002/qj.50.

——, and A. Shlyaeva, 2015: Scale-dependent background-error covariance localisation. *Tellus*, **67A**, 28027, https://doi.org/10.3402/tellusa.v67.28027.

——, J. Morneau, and C. Charette, 2013: Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. *Nonlinear Processes Geophys.*, **20**, 669–682, https://doi.org/10.5194/npg-20-669-2013.

——, and Coauthors, 2015: Implementation of deterministic weather forecasting systems based on ensemble–variational data assimilation at Environment Canada. Part I: The global system. *Mon. Wea. Rev.*, **143**, 2532–2559, https://doi.org/10.1175/MWR-D-14-00354.1.

Caron, J. F., and M. Buehner, 2018: Scale-dependent background error covariance localization: Evaluation in a global deterministic weather forecasting system. *Mon. Wea. Rev.*, **146**, 1367–1381, https://doi.org/10.1175/MWR-D-17-0369.1.

——, Y. Michel, T. Montmerle, and É. Arbogast, 2019: Improving background error covariances in a 3D ensemble-variational data assimilation system for regional NWP. *Mon. Wea. Rev.*, **147**, 135–151, https://doi.org/10.1175/MWR-D-18-0248.1.

Chen, J. H., and Coauthors, 2019: Advancements in hurricane prediction with NOAA's next-generation forecast system. *Geophys. Res. Lett.*, **46**, 4495–4501, https://doi.org/10.1029/2019GL082410.

Clayton, A. M., A. C. Lorenc, and D. M. Barker, 2013: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quart. J. Roy. Meteor. Soc.*, **139**, 1445–1461, https://doi.org/10.1002/qj.2054.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Derber, J., and A. Rosati, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.*, **19**, 1333–1347, https://doi.org/10.1175/1520-0485(1989)019<1333:AGODAS>2.0.CO;2.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, https://doi.org/10.1029/94JC00572.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Gasperoni, N. A., and X. Wang, 2015: Adaptive localization for the ensemble-based observation impact estimate using regression confidence factors. *Mon. Wea. Rev.*, **143**, 1981–2000, https://doi.org/10.1175/MWR-D-14-00272.1.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, 2011: Balance and ensemble Kalman filter localization techniques. *Mon. Wea. Rev.*, **139**, 511–522, https://doi.org/10.1175/2010MWR3328.1.

Gustafsson, N., J. Bojarova, and O. Vignes, 2014: A hybrid variational ensemble data assimilation for the High Resolution Limited Area Model (HIRLAM). *Nonlinear Processes Geophys.*, **21**, 303–323, https://doi.org/10.5194/npg-21-303-2014.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter–3D variational analysis scheme. *Mon. Wea. Rev.*, **128**, 2905–2919, https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

Holland, B., and X. Wang, 2013: Effects of sequential or simultaneous assimilation of observations and localization methods on the performance of the ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **139**, 758–770, https://doi.org/10.1002/qj.2006.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137, https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

——, and ——, 2005: Ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.*, **131**, 3269–3289, https://doi.org/10.1256/qj.05.135.

——, and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

Huang, B., and X. Wang, 2018: On the use of cost-effective Valid-Time-Shifting (VTS) method to increase ensemble size in the GFS hybrid 4DEnVar system. *Mon. Wea. Rev.*, **146**, 2973–2998, https://doi.org/10.1175/MWR-D-18-0009.1.

——, ——, and C. H. Bishop, 2019: The high-rank ensemble transform Kalman filter. *Mon. Wea. Rev.*, **147**, 3025–3043, https://doi.org/10.1175/MWR-D-18-0210.1.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126, https://doi.org/10.1016/j.physd.2006.11.008.

JCSDA, 2018: Data assimilation in the next generation global prediction system (NGGPS) Era: Initial implementation of FV3-based Global Forecast System (GFS). JCSDA, 34 pp., https://www.jcsda.org/s/2018_04JCSDAQuarterly.pdf.

Kleist, D. T., and K. Ide, 2015a: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

——, and ——, 2015b: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, https://doi.org/10.1175/MWR-D-13-00351.1.

——, D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP global

data assimilation system. *Wea. Forecasting*, **24**, 1691–1705, https://doi.org/10.1175/2009WAF2222201.1.

Kutty, G., and X. Wang, 2015: A comparison of the impacts of radiosonde and AMSU radiance observations in GSI based 3DEnsVar and 3DVar data assimilation systems for NCEP GFS. *Adv. Meteor.*, **2015**, 1–17, https://doi.org/10.1155/2015/280546.

Lei, L., and J. S. Whitaker, 2016: A four-dimensional incremental analysis update for the ensemble Kalman filter. *Mon. Wea. Rev.*, **144**, 2605–2621, https://doi.org/10.1175/MWR-D-15-0246.1.

——, and ——, 2017: Evaluating the trade-offs between ensemble size and ensemble resolution in an ensemble-variational data assimilation system. *J. Adv. Model. Earth Syst.*, **9**, 781–789, https://doi.org/10.1002/2016MS000864.

Lin, S.-J., 2004: A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP—A comparison with 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **129**, 3183–3203, https://doi.org/10.1256/qj.02.132.

——, 2017: Improving ensemble covariances in hybrid variational data assimilation without increasing ensemble size. *Quart. J. Roy. Meteor. Soc.*, **143**, 1062–1072, https://doi.org/10.1002/qj.2990.

——, N. E. Bowler, A. M. Clayton, S. R. Pring, and D. Fairbairn, 2015: Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Mon. Wea. Rev.*, **143**, 212–229, https://doi.org/10.1175/MWR-D-14-00195.1.

Lynch, P., and X.-Y. Huang, 1992: Initialization of the HIRLAM model using a digital filter. *Mon. Wea. Rev.*, **120**, 1019–1034, https://doi.org/10.1175/1520-0493(1992)120<1019:IOTHMU>2.0.CO;2.

Manly, B. F. J., 2006: *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Vol. 70, CRC Press, 480 pp.

Marchok, T., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., P1.13, http://ams.confex.com/ams/pdfpapers/37628.pdf.

Ménétrier, B., T. Montmerle, Y. Michel, and L. Berre, 2015: Linear filtering of sample covariances for ensemble-based data assimilation. Part I: Optimality criteria and application to variance filtering and covariance localization. *Mon. Wea. Rev.*, **143**, 1622–1643, https://doi.org/10.1175/MWR-D-14-00157.1.

Miyoshi, T., and K. Kondo, 2013: A multi-scale localization approach to an ensemble Kalman filter. *SOLA*, **9**, 170–173, https://doi.org/10.2151/sola.2013-038.

——, ——, and T. Imamura, 2014: The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Geophys. Res. Lett.*, **41**, 5264–5271, https://doi.org/10.1002/2014GL060863.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 44 pp., https://doi.org/10.21957/ps8gbwbdv.

Pan, Y., K. Zhu, M. Xue, X. Wang, M. Hu, S. G. Benjamin, S. S. Weygandt, and J. S. Whitaker, 2014: A GSI-based coupled EnSRF–En3DVar hybrid data assimilation system for the operational rapid refresh model: Tests at a reduced resolution. *Mon. Wea. Rev.*, **142**, 3756–3780, https://doi.org/10.1175/MWR-D-13-00242.1.

Putman, W., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, **227**, 55–78, https://doi.org/10.1016/j.jcp.2007.07.022.

Van Den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10-day forecast. *Wea. Forecasting*, **9**, 457–465, https://doi.org/10.1175/1520-0434(1994)009<0457:OTWFAE>2.0.CO;2.

Wang, X., 2010: Incorporating ensemble covariance in the grid-point statistical interpolation variational minimization: A mathematical framework. *Mon. Wea. Rev.*, **138**, 2990–2995, https://doi.org/10.1175/2010MWR3245.1.

——, 2011: Application of the WRF hybrid ETKF–3DVAR data assimilation system for hurricane track forecasts. *Wea. Forecasting*, **26**, 868–884, https://doi.org/10.1175/WAF-D-10-05058.1.

——, and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

——, and ——, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, https://doi.org/10.1256/qj.04.120.

——, and T. Lei, 2014: GSI-based four-dimensional ensemble–variational (4DEnsVar) data assimilation: Formulation and single-resolution experiments with real data for NCEP global forecast system. *Mon. Wea. Rev.*, **142**, 3303–3325, https://doi.org/10.1175/MWR-D-13-00303.1.

——, T. M. Hamill, J. S. Whitaker, and C. H. Bishop, 2007a: A comparison of hybrid ensemble transform Kalman filter–optimum interpolation and ensemble square root filter analysis schemes. *Mon. Wea. Rev.*, **135**, 1055–1076, https://doi.org/10.1175/MWR3307.1.

——, C. Snyder, and T. M. Hamill, 2007b: On the theoretical equivalence of differently proposed ensemble–3DVAR hybrid analysis schemes. *Mon. Wea. Rev.*, **135**, 222–227, https://doi.org/10.1175/MWR3282.1.

——, D. M. Barker, C. Snyder, and T. M. Hamill, 2008a: A hybrid ETKF–3DVAR data assimilation scheme for the WRF Model. Part I: Observation system simulation experiment. *Mon. Wea. Rev.*, **136**, 5116–5131, https://doi.org/10.1175/2008MWR2444.1.

——, ——, ——, and ——, 2008b: A hybrid ETKF–3DVAR data assimilation scheme for the WRF Model. Part II: Real observation experiments. *Mon. Wea. Rev.*, **136**, 5132–5147, https://doi.org/10.1175/2008MWR2445.1.

——, D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP global forecast system: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, https://doi.org/10.1175/MWR-D-12-00141.1.

——, H. G. Chipilski, C. H. Bishop, E. Satterfield, N. Baker, and J. S. Whitaker, 2021: A multiscale local gain form ensemble transform Kalman filter (MLGETKF). *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-20-0290.1, in press.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, https://doi.org/10.1175/MWR-D-11-00276.1.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, https://doi.org/10.1175/JAM2404.1.

Zapotocny, T. H., J. A. Jung, J. F. Le Marshall, R. E. Treadon, T. H. Zapotocny, J. A. Jung, J. F. Le Marshall, and R. E. Treadon, 2008: A two-season impact study of four satellite data types and rawinsonde data in the NCEP global data assimilation system. *Wea. Forecasting*, **23**, 80–100, https://doi.org/10.1175/2007WAF2007010.1.

Zhang, F., Y. Weng, J. A. Sippel, Z. Meng, and C. H. Bishop, 2009: Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble

Kalman filter. *Mon. Wea. Rev.*, **137**, 2105–2125, https://doi.org/10.1175/2009MWR2645.1.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the next generation global prediction system. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Zhu, Y., J. Derber, A. Collard, D. Dee, R. Treadon, G. Gayno, and J. A. Jung, 2014: Enhanced radiance bias correction in the National Centers for Environmental Prediction's gridpoint statistical interpolation data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **140**, 1479–1492, https://doi.org/10.1002/qj.2233.

Zong, H., and L. Wu, 2015: Synoptic-scale influences on tropical cyclone formation within the western North Pacific monsoon trough. *Mon. Wea. Rev.*, **143**, 3421–3433, https://doi.org/10.1175/MWR-D-14-00321.1.