A Multiresolution Ensemble Hybrid 4DEnVar with Variable Ensemble Sizes to Improve Global and Tropical Cyclone Track Numerical Prediction

ERIN A. JONES[a] AND XUGUANG WANG[a]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 11 July 2022, in final form 23 January 2023)

ABSTRACT: The current global operational four-dimensional ensemble-variational (4DEnVar) data assimilation (DA) system at NCEP adopts a background ensemble at a reduced resolution, which restricts the range of spatial scales that the ensemble background error covariance can resolve. A prior study developed a multiresolution ensemble 4DEnVar method and determined that this approach can provide a comparable forecast to an approach using solely high-resolution members, while substantially reducing the computational cost. This study further develops the multiresolution ensemble 4DEnVar approach to allow for a flexible number of low- and high-resolution ensemble members as well as varying localization length scales between the high- and low-resolution ensembles. Three 4DEnVar experiments with the same computational costs are compared. The first experiment has an 80-member high-resolution background ensemble with single-scale optimally tuned localization (SR-High). The second and third experiments utilize the multiresolution background ensembles. One has 130 low-resolution and 40 high-resolution members (MR170) while the other has 180 low-resolution members and 24 high-resolution members (MR204). Both multiresolution ensemble experiments utilize differing localization radii with ensemble resolution. Despite having the same costs, both MR170 and MR204 improves global forecasts and decreases tropical cyclone track errors for up to 5 days' lead time compared to SR-High. Improvements are most apparent in larger-scale features, such as jet streams and the environmental steering flow of tropical cyclones. Additionally, MR170 outperforms MR204 in terms of global and tropical cyclone track forecasts, demonstrating the value of both increasing sampling at large scales and retaining substantial information at small scales.

KEYWORDS: Hurricanes/typhoons; Data assimilation; Numerical weather prediction/forecasting

1. Introduction

There has been steady improvement in global NWP from the latter half of the twentieth century through the present (e.g., Hamill et al. 2010; Magnusson and Källén 2013; Bauer et al. 2015; Benjamin et al. 2019). These improvements have stemmed from advancements in computing resources, model dynamics, model physics, observing systems, and data assimilation. Recent studies have highlighted the particular importance of improving the initial conditions (ICs) of numerical models through the advancement of data assimilation (DA) methods (Magnusson et al. 2019; Chen et al. 2019).

Operational numerical prediction centers utilize various advanced DA techniques. One popular category includes the ensemble-based DA approaches, such as various forms of EnKF (e.g., Evensen 1994; Houtekamer and Mitchell 2001; Bishop et al. 2001; Anderson 2001; Whitaker and Hamill 2002; Wang and Bishop 2003; Wang et al. 2004; Hunt et al. 2007) and hybrid ensemble–variational (EnVar; e.g., Hamill and Snyder 2000; Lorenc 2003; Buehner 2005; Wang et al. 2007, 2008a,b; Wang 2010; Wang et al. 2013; Wang and Lei 2014) methods. The current global operational system at NCEP uses a hybrid 4DEnVar technique with a single control member at high resolution and 80 ensemble members at a reduced resolution (JCSDA 2018). Limitations in computational resources require examining the optimal balance between model resolution and ensemble size in the context of ensemble-based DA. The use of a high-resolution ensemble resolves smaller-scale background errors (e.g., Hamill and Whitaker 2005). However, increasing the ensemble size reduces the sampling error of the ensemble-estimated background error covariances (BECs; e.g., Miyoshi et al. 2014; Huang and Wang 2018).

Lei and Whitaker (2017) used a given computational cost to examine either increasing ensemble size or ensemble resolution in 4DEnVar. They found that increasing ensemble resolution was more advantageous than increasing ensemble size due to the further reduction of errors at small scales. When computational cost was not constrained, Houtekamer et al. (2014) and Hamrud et al. (2015) found benefit in increasing both ensemble size and ensemble resolution when using EnKF methods. These studies used the approach of universally increasing resolution of all background ensemble members or directly increasing the size of the single-resolution background ensemble (SR-ENS).

Studies have also explored the optimal ensemble size and resolution for ensemble-based DA using a multiresolution background ensemble approach (MR-ENS, hereafter), where some background ensemble members are at a low resolution and others are at a high resolution. Rainwater and Hunt (2013) examined the MR-ENS approach in the local ensemble transform Kalman filter (LETKF) using an idealized Lorenz model (Lorenz 1996). At a similar computational cost for the forecast, they found that the MR-ENS LETKF outperformed a SR-ENS LETKF. The study, however, was limited by providing comparisons with a constraint only on the forecast computational time, neglecting the computational cost for the

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

DOI: 10.1175/MWR-D-22-0186.1

DA steps. Additionally, since the tests were conducted in a Lorenz model framework, their experimental context was highly simplified compared to the operational NWP.

Kay and Wang (2020) introduced a MR-ENS approach in the Gridpoint Statistical Interpolation analysis system (GSI) 4DEnVar. They determined that the MR-ENS 4DEnVar with 40 high-resolution members and 40 low-resolution members improved upon the analysis and forecast of a SR-ENS 4DEnVar with 80 low-resolution members. Additionally, forecast errors of the MR-ENS approach were comparable to a SR-ENS 4DEnVar with 80 high-resolution members while MR-ENS reduced the computational cost by 40%. However, this study had several limitations. First, they utilized the same number of low- and high-resolution members in the MR-ENS background ensemble, which inhibited the exploration of the effect of diverse sampling at different scales of interest. Second, they used the same covariance localization distance for the low- and high-resolution ensemble in the MR-ENS configuration, which constrained the effectiveness of estimating error covariances at multiple scales. Third, their comparison did not constrain the computation cost, which limited the ability to directly compare improvements.

This study aims to further develop and explore the MR-ENS approach in 4DEnVar. First, this study will advance the MR-ENS methodology of Kay and Wang (2020) to allow the use of different ensemble sizes for differing resolution in the background ensemble. For example, a larger ensemble size can be assigned for the low-resolution ensemble members than the high-resolution members and vice versa. Therefore, the new capability will allow for flexible sampling of the scales of interest. Second, the MR-ENS approach is further developed to allow the high- and low-resolution members to use different covariance localization radii. For example, the high-resolution members use a tighter localization than the low-resolution members. This capability will allow the high-resolution members to focus on estimating the small-scale error covariance and the low-resolution members on the large-scale error covariance. Third, the MR-ENS experiments are configured to ensure the same computational cost, including both the DA and forecast computational times, as the 4DEnVar experiment with an ensemble solely composed of high-resolution members (SR-High, hereafter). In this experiment design, the total number of ensemble members in MR-ENS that can resolve large-scale background errors is increased compared with SR-High. Correspondingly, the total number of ensemble members in MR-ENS that can resolve small scales is decreased relative to SR-High. Using this experiment setup, this paper addresses the following questions: how does the use of more low-resolution background ensemble members and fewer high-resolution background ensemble members impact the global analysis and subsequent global forecast? In which regions and associated with which meteorological features are the greatest impacts occurring? How does the impact on the analysis and forecast change for varying spatial scales? MR-ENS 4DEnVar is one of the approaches toward achieving an effective multiscale DA system for the next-generation global NWP (e.g., Wang et al. 2021). This study focuses on evaluating and diagnosing deterministic analyses and forecasts.

The rest of this paper is organized as follows. Section 2 describes the formulation of the newly extended MR-ENS hybrid 4DEnVar. Section 3 describes configurations of the SR-High and MR-ENS experiments, including DA and model parameters. Results comparing the MR-ENS 4DEnVar and SR-High 4DEnVar are discussed in section 4. The main conclusions are summarized in section 5.

2. Formulation of the multiresolution ensemble hybrid 4DEnVar

The SR-High experiment in this study utilizes the hybrid 4DEnVar formulation as described in Wang and Lei (2014). The MR-ENS 4DEnVar method as described in Kay and Wang (2020) extends the hybrid 4DEnVar formulation to incorporate BECs from a multiresolution ensemble. The variables representing the low- and high-resolution ensembles are represented by L and H, respectively. The total number of state variables at low and high resolution are denoted by N_L and N_H, respectively. For this formulation, the analysis increments $\mathbf{x}'_t$, a column matrix of size N_H, are created for the control member at times $t = 1, \ldots, L$ and are defined by

$$\mathbf{x}'_t = \mathbf{x}'_1 + \mathbf{u}\sum_{k=1}^{K_L}[(\boldsymbol{\alpha}_L)_k \circ (\mathbf{x}_L)_{k,t}] + \sum_{k=1}^{K_H}[(\boldsymbol{\alpha}_H)_k \circ (\mathbf{x}_H)_{k,t}], \quad (1)$$

where $\mathbf{x}'_1$ is the analysis increment associated with the static error covariance. The linear transform matrix $\mathbf{u}$ of size $N_H \times N_L$ interpolates low-resolution grids to high-resolution. $(\boldsymbol{\alpha}_L)_k$ and $(\boldsymbol{\alpha}_H)_k$ are the kth augmented control vectors (Wang 2010) for the low- and high-resolution ensembles with sizes N_L and N_H, respectively. The symbol "$\circ$" represents the Schur product. Different from Kay and Wang (2020), the numbers of ensemble members for the low-resolution (K_L) and high-resolution (K_H) may have different values. $(\mathbf{x}_L)_{k,t}$ and $(\mathbf{x}_H)_{k,t}$ are the kth background ensemble perturbations with sizes N_L and N_H at time t normalized by $\sqrt{K_L - 1}$ and $\sqrt{K_H - 1}$.

The analysis increments are found by the minimization of the following cost function J with respect to the augmented control vectors $\boldsymbol{\alpha}_L$ and $\boldsymbol{\alpha}_H$:

$$\begin{aligned} J[\mathbf{x}'_1, \boldsymbol{\alpha}_L, \boldsymbol{\alpha}_H] &= \beta_1 J_1 + \beta_L J_L + \beta_H J_H + J_o \\ &= \frac{1}{2}\beta_1(\mathbf{x}'_1)^{\mathrm{T}}\mathbf{B}_1^{-1}(\mathbf{x}'_1) + \frac{1}{2}\beta_L(\boldsymbol{\alpha}_L)^{\mathrm{T}}\mathbf{A}_L^{-1}(\boldsymbol{\alpha}_L) \\ &\quad + \frac{1}{2}\beta_H(\boldsymbol{\alpha}_H)^{\mathrm{T}}\mathbf{A}_H^{-1}(\boldsymbol{\alpha}_H) \\ &\quad + \frac{1}{2}\sum_{t=1}^{L}(\mathbf{H}\mathbf{x}'_t - \mathbf{y}_t^o)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x}'_t - \mathbf{y}_t^o). \end{aligned} \quad (2)$$

The static BEC $\mathbf{B}_1$, as traditionally used in GSI 3DVar (Wu et al. 2002), is associated with the analysis increment $\mathbf{x}'_1$; $\boldsymbol{\alpha}_L$ and $\boldsymbol{\alpha}_H$ are column matrices that represent the concatenation of augmented control vectors for the low- and high-resolution ensembles with sizes $N_L K_L$ and $N_H K_H$, respectively; and $\mathbf{A}_L$ and $\mathbf{A}_H$, with sizes $N_L K_L \times N_H K_H$ and $N_L K_L \times N_H K_H$, are used to determine the covariance localization for the low- and high-resolution ensembles, respectively. Matrices $\mathbf{A}_L$ and $\mathbf{A}_H$

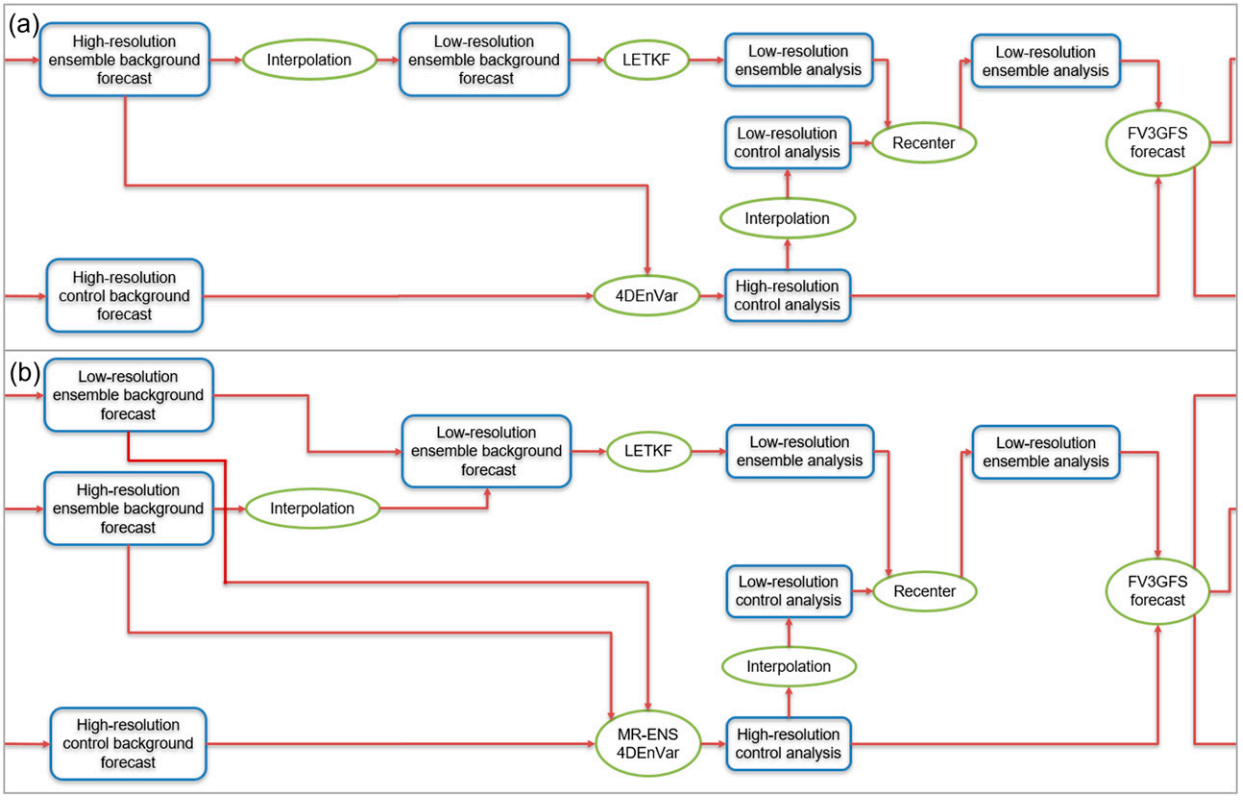

FIG . 1. Flow charts for a cycle of the (a) SR-High and (b) MR-ENS experiments.

consist of both horizontal and vertical components. A spectral filter transform is employed for the horizontal localization (e.g., Wang and Lei 2014; Kleist and Ide 2015), where a Gaspari and Cohn (1999) function with set *e*-folding distances vary with the vertical level of the model (e.g., Kleist and Ide 2015; Lei and Whitaker 2017). A recursive filter transform (Hayden and Purser 1995) is applied for vertical covariance localization and is given as a level-invariant length scale. Different from Kay and Wang (2020), the covariance localization scales may also vary between the low- and high-resolution ensembles. The terms β_1, β_L, and β_H are the weights given to the static, low-resolution ensemble, and high-resolution ensemble BECs and are constrained by $1/\beta_1 + 1/\beta_L + 1/\beta_H = 1$, following Wang et al. (2007) and Kay and Wang (2020). These weights are scalars which currently have no scale-dependence. Other terms in Eq. (2) are as follows: **H** is the linearized observation operator; $\mathbf{y}_t^o$ is the innovation vector at time *t*; **R** is the observation error covariance matrix.

3. Experimental design

The GSI-based 4DEnVar DA system was cycled every 6 h for a 5-week period—from 1800 UTC 25 August to 1200 UTC 29 September 2017. The background and 120-h forecasts were generated by the GFS using the Finite Volume Cubed-Sphere dynamical core (FV3GFS; Harris et al. 2021). In all statistical verification, the first week was removed to account for spinup time. The system assimilates the NCEP operational global conventional and satellite observations[1] over a 6-h DA window.

a. Single high-resolution (SR-High) experiment

A flowchart of the general steps in the first experiment, similar to Wang et al. (2013, their Fig. 1b), is given in Fig. 1a. In the SR-High experiment, the background control member and the background ensemble were at the same resolution (T766 or ~25 km). It is noted that although SR-High is designed as a baseline for this study, it is different from the operational configuration. Additionally, the resolution of the SR-High and MR-ENS experiments in this study is coarser compared to that of the operational due to limited computing resources. To distinguish from the second experiment where some background ensemble members were run at lower resolution compared to the control background forecast, we name the first experiment SR-High, standing for "single high-resolution experiment." In SR-High, as shown in Fig. 1a, the control analysis was performed by 4DEnVar. Its increments associated with the static and ensemble BECs were, by construction, both at the same high resolution. As in the current operational 4DEnVar system, 80 ensemble members were ingested. The static and ensemble BECs were given weights of 12.5% and

[1] https://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/table_2.htm and https://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/table_18.htm.

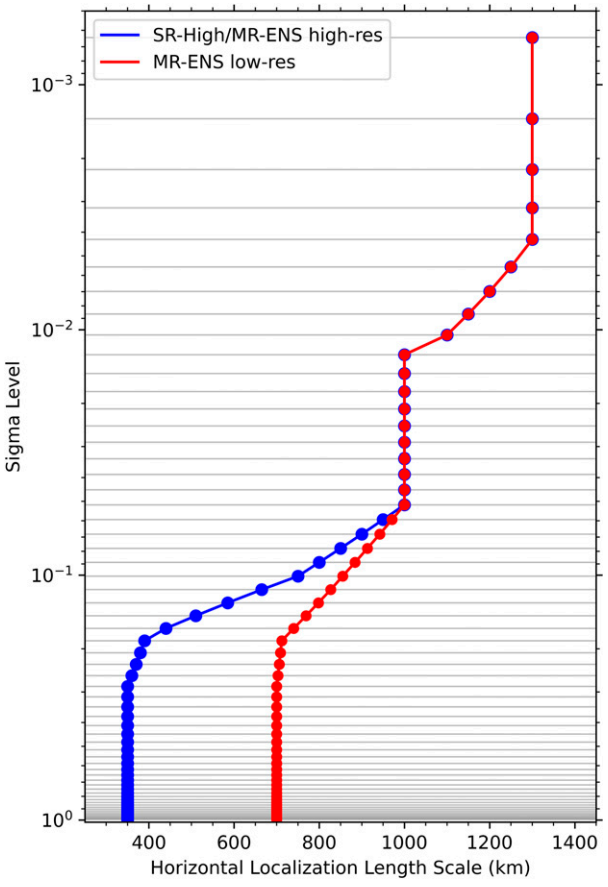

FIG. 2. Level-dependent localization length scales in *e*-folding distance (km) for 4DEnVar for the SR-High experiment and high-resolution ensemble for the MR-ENS experiments (blue) and for the low-resolution ensemble for the MR-ENS experiments (red).

87.5%, respectively (Lei and Whitaker 2017; Huang and Wang 2018; Kay and Wang 2020; Huang et al. 2021). As shown in the blue line in Fig. 2, level-dependent horizontal covariance localization was utilized. The horizontal *e*-folding distance was set to 350 km from the surface to ~300 hPa (hl1), 1000 km between approximately 56 and 14 hPa (hl2), and 1300 km from around 5 hPa to the model top (hl3) with a linear transition between these levels. In the vertical direction, a constant localization *e*-folding distance of 0.5 scale heights was applied. These localization distances have been identified in this study through performing sensitivity tests. The same values are utilized operationally with the FV3GFSv15 4DEnVar and in studies such as Huang et al. (2021).

For the ensemble update component, 4D LETKF was adopted (Bishop et al. 2001; Wang and Bishop 2003; Wang et al. 2004; Hunt et al. 2007; Lei et al. 2018). Multiple localization distances were tested for 4D LETKF in this study. The level-dependent Gaspari and Cohn (1999) cutoff distance for horizontal and vertical covariance localization, same as those used in Lei and Whitaker (2017), were found optimal and used in this study. The horizontal cutoff distances are equivalent to *e*-folding distances of 263 km for hl1, 1000 km for hl2, and 1300 km for hl3 were used. The constant vertical localization cutoff distance, equivalent to 0.375 scale heights, was adopted. Following Kay and Wang (2020), due to computational constraints, the high-resolution background ensemble was interpolated to low resolution (T382 or ~50 km) before the LETKF update. Multiplicative inflation (Whitaker and Hamill 2012) was used to increase ensemble spread by relaxing the spread of the low-resolution posterior ensemble to 85% of the spread of the low-resolution background ensemble (e.g., Lei and Whitaker 2017; Kay and Wang 2020; Huang et al. 2021). The updated 4D LETKF ensemble analyses were recentered on an interpolated (T766 to T382) 4DEnVar control analysis.

The background forecasts for the single control member and 80-member ensemble were initialized by the 4DEnVar control analysis and the recentered 4D LETKF ensemble analyses, respectively. These forecasts were run at high resolution with 64 vertical model levels by the FV3GFS. To reduce imbalance in the forecast, a four-dimensional incremental analysis update (4DIAU; Bloom et al. 1996; Lorenc et al. 2015) was used in both the control and ensemble forecasts. High resolution is considered to be C384, or ~25 km. Low resolution is C192, or ~50 km. The configuration of model-specific parameters is detailed in Zhou et al. (2019).

b. Multiresolution ensemble (MR-ENS) experiments

The flowchart for the MR-ENS experiments is shown in Fig. 1b. The MR-ENS 4DEnVar experiments utilized the methodology from section 2, where the high-resolution control member was updated using a high-resolution static BEC, and low-resolution (T382 or ~50 km) and high-resolution (T766 or ~25 km) ensemble BECs. As discussed in section 2, the expanded approach allows for varying ensemble sizes. Two MR-ENS configurations that had computational costs nearly equal to SR-High (Table 1), were examined. The first configuration used 130 low-resolution members and 40 high-resolution members for a total ensemble size of 170 members and will thus be denoted as MR170. The second configuration utilized a larger number of low-resolution members, at 180 members, but a smaller number of high-resolution members, at 24 members, for a total of 204 members and, therefore, will be denoted as MR204. These two experiments were designed to test the sensitivity of the analyses and forecasts to the balance between the large-scale and small-scale sampling. The MR-ENS configuration that led to the smallest forecast errors was used in additional comparisons to the SR-High experiment to determine what mechanisms were driving the differences between the experiments.

For the 4DEnVar step, as in Kay and Wang (2020), the static BEC was given a weight of 12.5% and the remaining weight was split evenly between the low- and high-resolution ensemble BECs at 43.75% each. Given equal weights are assigned to the low- and high-resolution ensembles, compared to SR-High, effectively smaller weight is given to small-scale BECs. The vertical localization was applied, as in the SR-High experiment, with a constant localization *e*-folding distance of 0.5 scale heights. As described in section 2, the expanded MR-ENS approach allows

TABLE 1. The computational cost for the SR-High, MR170, and MR204 experiments for each DA and forecast step. Ratios are calculated in reference to the SR-High experiment. Costs of all experiments were estimated using the vjet partition on NOAA's Research and Development High-Performance Computing System Jet machine. I/O cost is not included in the estimation. Since I/O costs can vary greatly based on frequency of output, number of variables outputted, and if any parallel I/O techniques (Balle and Johnsen 2016) are used, for this comparison, I/O costs were not included. Core hours are highlighted in boldface.

	Experiment		
	SR-High	MR170	MR204
4DEnVar			
No. of cores	1260	1260	1260
Time (h)	0.92	0.91	0.91
Core hours	**1159.20**	**1146.60**	**1146.60**
LETKF			
No. of cores	1260	1260	1260
Time (h)	0.21	0.67	0.90
Core hours	**264.60**	**844.20**	**1134.00**
Control member forecast			
No. of cores	396	396	396
Time (h)	0.079	0.079	0.079
Core hours	**31.28**	**31.28**	**31.28**
High-resolution ensemble forecast			
No. of cores	396	396	396
Time (h)	0.079	0.079	0.079
Ensemble size	80	40	24
Core hours	**2502.72**	**1251.36**	**750.82**
Low-resolution ensemble forecast			
No. of cores		108	108
Time (h)		0.047	0.047
Ensemble size		130	180
Core hours	**0.00**	**659.88**	**913.68**
Total core hours	**3957.80**	**3933.32**	**3976.38**
Ratio	1.000	0.994	1.005

for varying horizontal localization distances between the low- and high-resolution ensembles. Prior studies have indicated optimal localization distances increase with increasing ensemble size, leading to greater dynamical balance and allowing for realistic larger-scale correlations to be maintained (e.g., Houtekamer and Mitchell 2001; Hamill et al. 2001; Ying et al. 2018; Wang et al. 2021). For the high-resolution ensemble in these experiments, the level-dependent localization distances (hl1, hl2, and hl3) were equal to those described above in the SR-High experiment (blue line in Fig. 2). For the low-resolution ensemble, which has a much larger ensemble size, hl1 was set to an *e*-folding distance of 700 km (red line in Fig. 2) as in Kleist and Ide (2015). Their parameters were determined through sensitivity tests.

Similar to the SR-High experiment and Kay and Wang (2020), given the computational constraint, the 4D LETKF updates to the ensemble were made at low-resolution by interpolating the high-resolution ensemble members from T766 to T382. Though some small-scale information is lost in the interpolation for the 4D LETKF step for both SR-High and MR-ENS, their ensemble resolution differences were maintained in both the background ensemble forecast step and in the 4DEnVar steps. Testing indicates that, by the 3-h forecast, running the forecast model at high resolution recovers greater than 97% of the loss of such small-scale information. By the 9-h forecast, the recovery is greater than 99%. Based on the sensitivity tests and past studies (e.g., Lei and Whitaker 2017), 4D LETKF localization and inflation parameters used for MR-ENS adopted were the same as SR-High, described in section 3a. The FV3GFS background forecasts were completed at high-resolution for the control member and high-resolution ensemble and at low-resolution for the low-resolution ensemble. 4DIAU is applied similarly to SR-High, except the analysis increments associated with the low-resolution ensemble are mapped onto a low-resolution model grid.

4. Results

a. Multiscale analysis

To determine the multiscale impact of the extended MR-ENS and single-resolution ensemble 4DEnVar on the analysis, analysis increments were calculated, decomposed

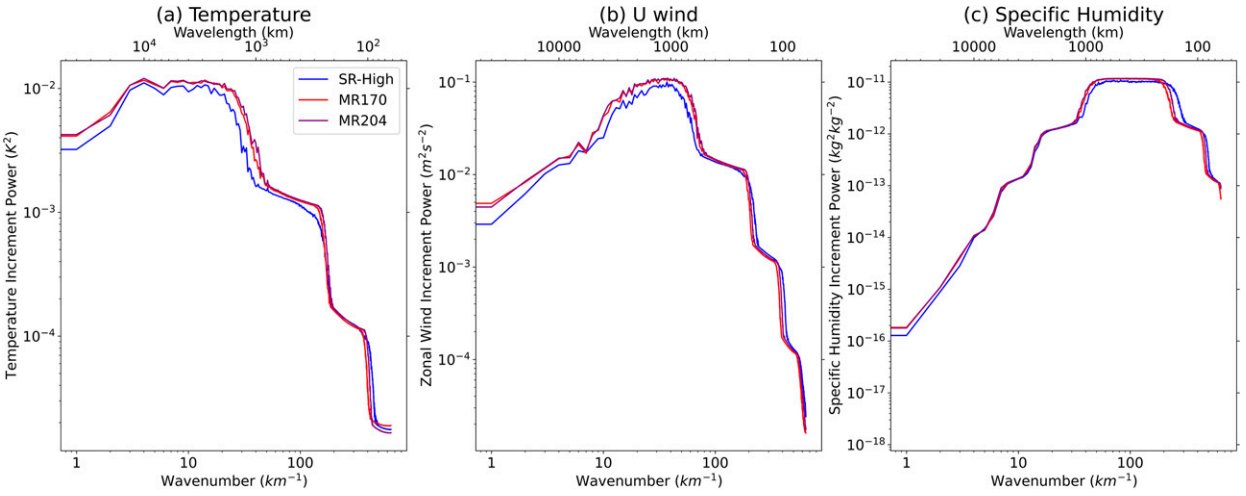

FIG. 3. Averaged analysis increment power at 500 hPa over all DA cycles as a function of wavenumber and wavelength for (a) temperature (K^2), (b) zonal wind speed ($m^2\ s^{-2}$), and (c) specific humidity ($kg^2\ kg^{-2}$) for the SR-High (blue), MR170 (red), and MR204 (purple) experiments.

into spectral space using spherical harmonics, and averaged over all cycles. The 500-hPa analysis increment power spectra for temperature, zonal wind, and specific humidity are shown in Fig. 3. For temperature and zonal wind, MR170 has larger increments for larger atmospheric scales (wavenumbers $<$ ~150 or wavelengths $>$ ~200 km) when compared with SR-High. Such an increase of increments at large scales is reduced for specific humidity. For smaller scales (wavenumbers $>$ ~150 or wavelengths $<$ ~200 km), the MR170 experiment has similar or slightly smaller analysis increments.

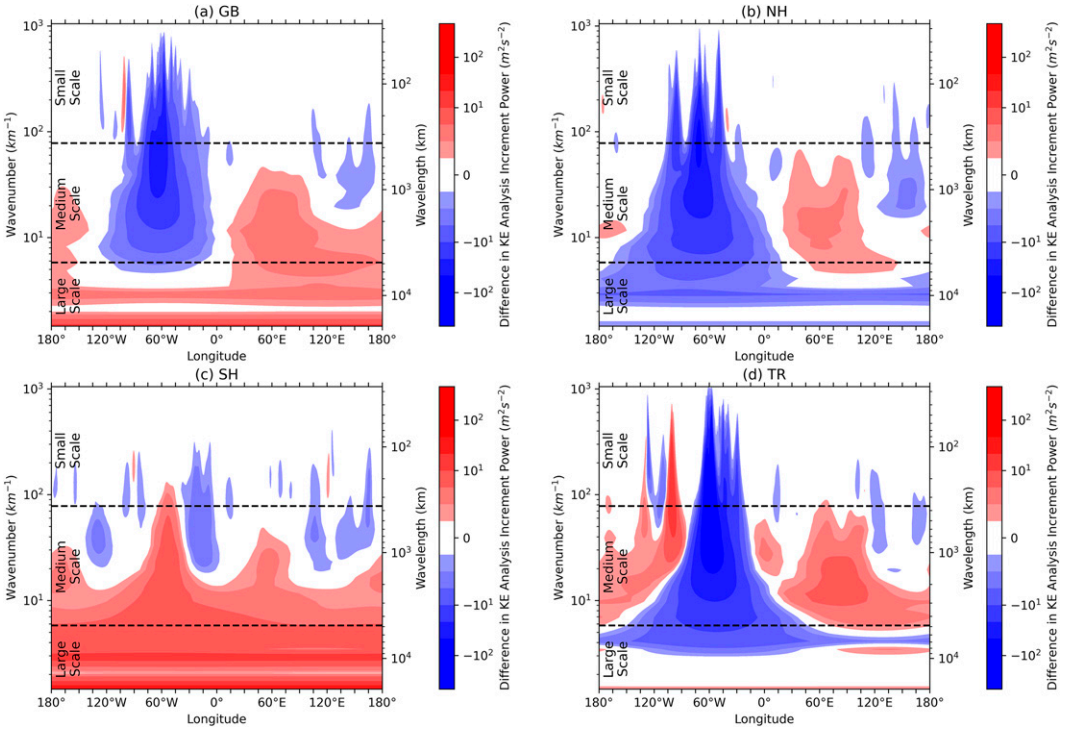

FIG. 4. Difference in horizontal kinetic energy wavelet analysis between MR170 and SR-High averaged over all DA cycles and over the (a) global, (b) Northern Hemisphere extratropical, (c) Southern Hemisphere extratropical, and (d) tropical latitudes. The x axis represents the longitude. The left and right y axes represent the wavenumber and corresponding wavelength, respectively. Red and blue shading represent scales and longitudes where MR170 and SR-High, respectively, have larger horizontal kinetic energy analysis increments. The cutoffs between small, medium, and large scales are denoted by dashed black lines.

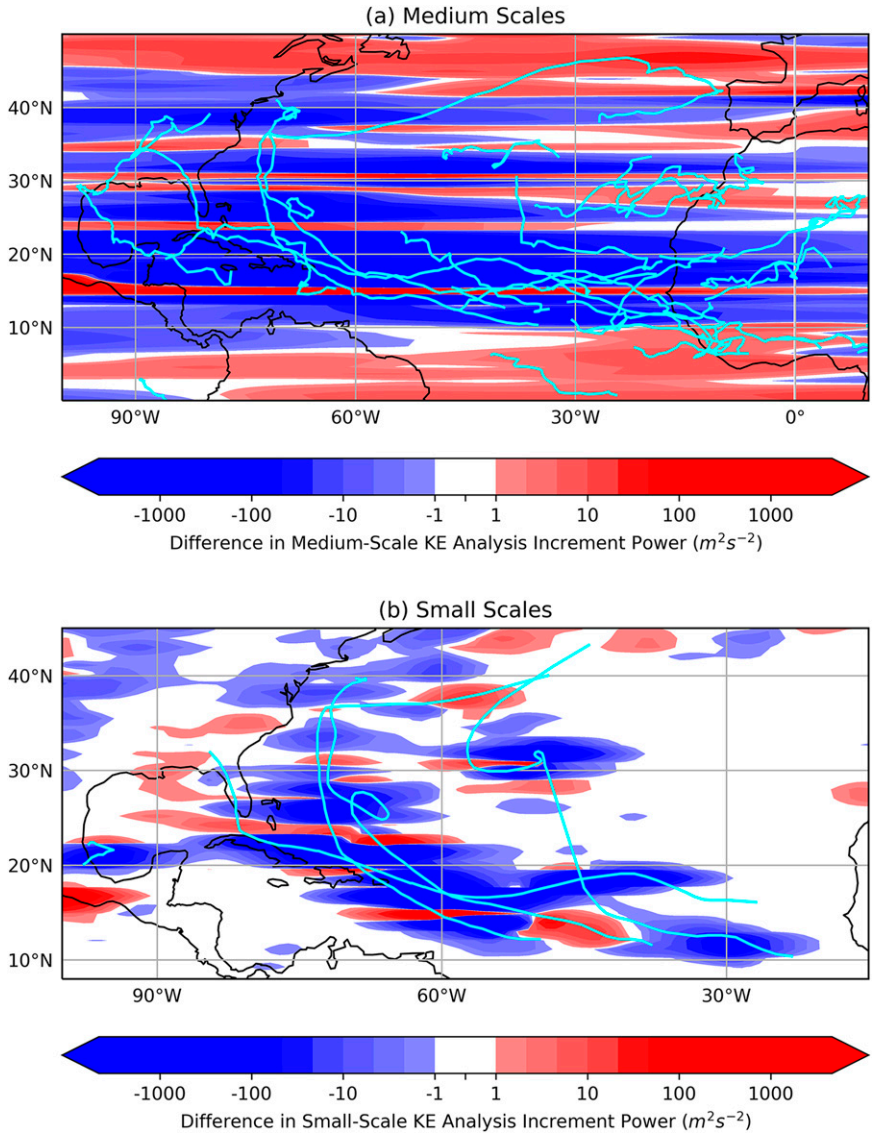

FIG. 5. Difference in the wavelet horizontal kinetic energy analysis increment power ($m^2\ s^{-2}$) between MR170 and SR-High averaged over all cycles and over (a) medium scales (wavelet scales 16–30) and (b) small scales (wavelet scales 1–15). Red and blue shading represent latitudes and longitudes where MR170 and SR-High, respectively, have larger horizontal kinetic energy analysis increments. The tracks of (a) the tropical easterly waves that occurred at 850 hPa (Hollis 2021) or (b) the Atlantic hurricanes that occurred during September 2017 are plotted in cyan.

These results are consistent with the MR-ENS and SR-High methodologies. MR170 has more ensemble members to resolve the large-scale error covariances than SR-High. Additionally, MR170 applies a broader localization scale to its low-resolution ensemble BECs, while SR-High applies a shorter localization for its entire single-resolution ensemble BECs. Such differences allow MR-ENS to make larger adjustments at large scales. For variables which have generically smaller scales, such an impact on the large-scale increment is limited. For smaller scales, SR-High has more high-resolution members to resolve the small-scale error covariance than MR170. The high-resolution ensemble covariance is also assigned a higher weight than MR170. Such differences lead to slightly larger analysis increments at smaller scales by SR-High. As discussed in section 5, combining SR-High and MR-ENS with multiscale DA techniques, such as scale-dependent localization (SDL; Buehner and Shlyaeva 2015; Huang et al. 2021), may likely vary the analysis increment power differences between the two experiments. Such topics will be left for future studies.

A spatial wavelet diagnostic on the analysis increments is also performed (Torrence and Compo 1998). A wavelet analysis can diagnose both the location and spatial scale of decomposed analysis increments. The horizontal kinetic energy analysis

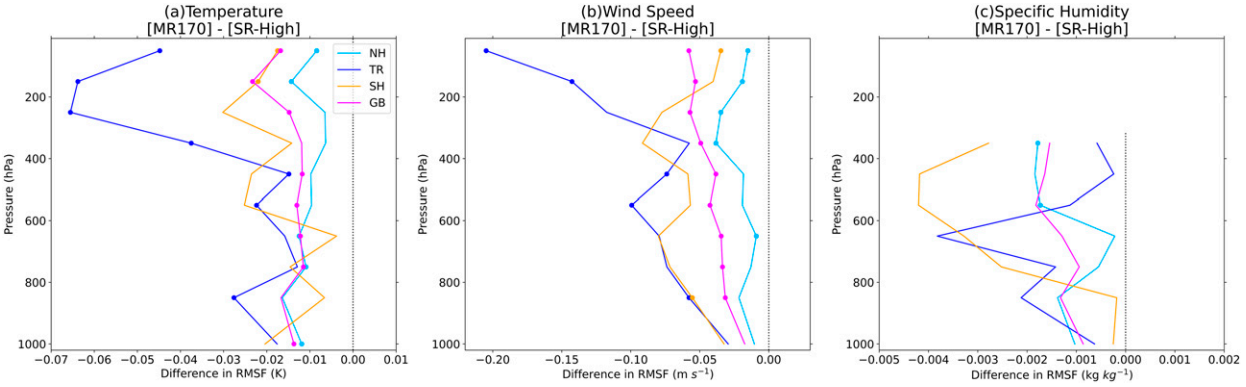

FIG. 6. Difference in the root-mean-square fit (RMSF) of the 6-h forecast of (a) temperature (K), (b) wind speed (m s^{-1}), and (c) specific humidity (kg kg^{-1}) to the rawinsonde observations between the MR170 and SR-High experiments as a function of pressure, averaged over all cycles every 100 hPa over global (magenta), Northern Hemisphere extratropical (light blue), Southern Hemisphere extratropical (orange), and tropical (dark blue) latitudes. Dots indicate levels where the RMSF difference is statistically significant using a paired *t* test and a 95% confidence interval.

increments for each analysis time were broken up into Morlet wavelets (Domingues et al. 2005) for each latitude and vertical level. The wavelets were averaged over all vertical levels, all cycles, and latitudes in four regions: globally (GB; all latitudes), the Northern Hemisphere extratropics (NH; northward of 20°N), the Southern Hemisphere extratropics (SH; southward of 20°S), and the tropics (TR; between 20°N and 20°S). The difference in these averages for the MR170 and SR-High experiments are shown in Fig. 4. The scales chosen for the analysis were from 2 to $2^{(n/4)}$ grid points, where n is the total number of scales, or 39 in this case. In Fig. 4, the y axes have been converted to wavenumber and wavelength—analogous to the x axes in the above Fourier spectral analyses—for aid in interpreting what scales are affected.

For the largest scales, there is a general increase in kinetic energy analysis increments for the MR170 experiment averaged over all latitudes (Fig. 4a). This increase is mainly dominated by the SH region (Fig. 4c), indicating that at the largest scales, MR170 has a greater impact on the analysis increments in the SH region. At the smallest scales, averaged over all latitudes, SR-High has substantially greater analysis increment power between 15° and 90°W, dominated by TR and NH latitudes (Figs. 4b,d). At medium scales, MR170 and SR-High produce mixed results with regards to their effect on the kinetic energy analysis increments. For most longitudes, MR170 still has greater analysis increment power. Between about 120°W and 0°, the global analysis increment power for SR-High is substantially increased. As shown in Figs. 4b and 4d, this increase is dominated by NH and TR in longitudes associated with the Atlantic Ocean, the Caribbean Sea, the Gulf of Mexico, and the U.S. East Coast. Examining the vertical profile, at upper levels, MR170 generally has greater horizontal kinetic energy analysis increment power, especially at medium to large scales, and shifts to lower power at nearly all scales at lower levels (not shown).

To determine what physical processes are associated with the greater medium-scale increment of SR-High between 120°W and 0°, instead of averaging the wavelets over latitude, as in Fig. 4, the wavelets were averaged over small- (wavelet scales of 1–15 or wavelengths $<$ ~350 km) and medium-scales (wavelet scales of 16–30 or wavelengths between approximately 350 and 5000 km). The 350-km cutoff between small and medium scales was chosen, consistent with the localization radius for the high-resolution ensembles. These wavelengths capture the most intense impacts from tropical cyclones (TCs) and tropical easterly waves (e.g., Carlson 1969; Burpee 1972; Engel et al. 2017). The cutoff between medium and large scales was chosen to include all impacts from these systems. For example, the maximum wavelength of a tropical easterly wave is ~5000 km (e.g., Reed et al. 1977). As shown in Fig. 5a, the greater analysis increment power of the SR-High analysis at medium scales is primarily collocated with the high density of tropical easterly waves in the Atlantic. Additionally, the greater analysis increment power of the SR-High analysis at small scales is located at the northern tropical Atlantic and near the U.S. East Coast (Fig. 5b). Four major hurricanes (Irma, Jose, Lee, and Maria) fall within that area during the experiment period. Bands of red within the TC basin are dominated by storms undergoing rapid changes in intensity, indicating the role of multiple scales and their interactions for TC rapid intensity changes.

In summary, SR-High tends to make larger increments in relatively small scales due to the greater number of high-resolution ensemble members and their larger weight. In more synoptically forced regions, however, as is the case in the austral late winter and early spring during which this study occurs, MR170 makes larger corrections in relatively larger scales, as would be expected with a larger number of low-resolution members and a broader localization distance applied to these members. It is worth noting that large analysis increments do not necessarily indicate a better analysis or subsequent forecast a priori; they instead indicate that the observations adjusted the analysis further away from the background forecast. Subsequent sections will examine the impact of these corrections on the global and regional analysis and forecast errors.

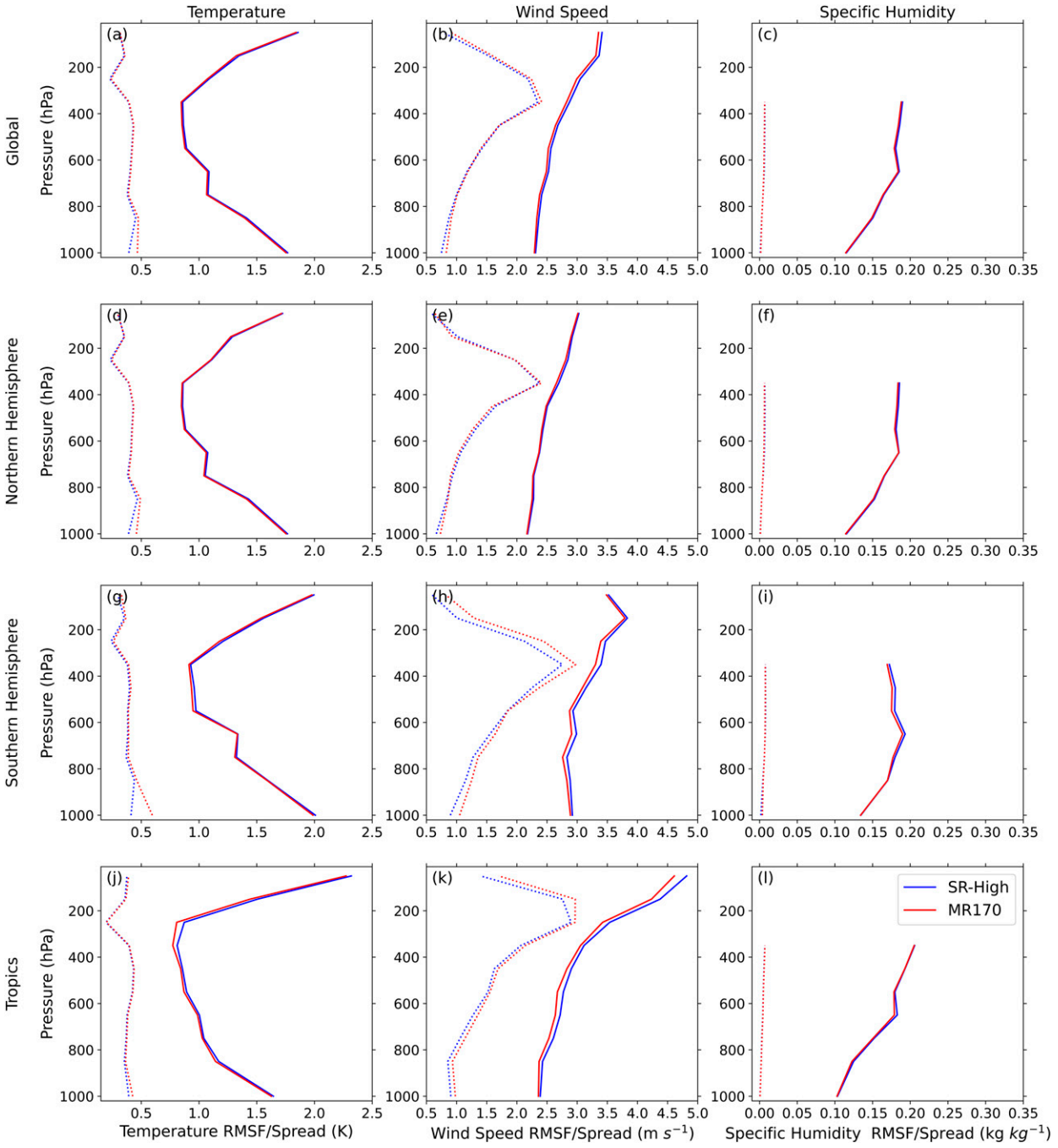

FIG. 7. RMSF of the 6-h forecast to the rawinsonde observations (solid lines) and square root of the sum of the 6-h background ensemble variance and the observation error variance (dotted lines) for the (left) temperature (K), (center) wind speed (m s^{-1}), and (right) specific humidity (kg kg^{-1}) for the MR170 (red) and SR-High (blue) experiments as a function of pressure averaged over all cycles every 100 hPa over (a)–(c) global, (d)–(f) Northern Hemisphere extratropical, (g)–(i) Southern Hemisphere extratropical, and (j)–(l) tropical latitudes.

b. Short-term forecast errors and ensemble spread

To determine if the greater analysis increments translate into a more accurate forecast, the difference between the RMS fit (RMSF) of the 6-h background forecast to the observations for all cycles was calculated and averaged over vertical levels in increments of 100 hPa. The RMSF was broken down into the same four regions as described above: GB, NH, SH, and TR. In Fig. 6, the difference in RMSF between the two experiments as well as the statistical significance, which was calculated using a paired *t* test (e.g., Walpole et al. 2011) with

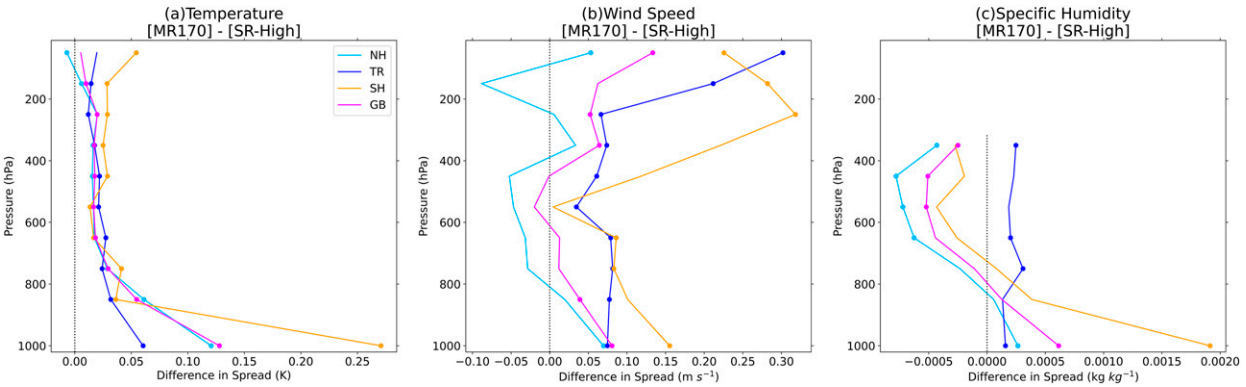

FIG. 8. As in Fig. 6, but for difference of the square root of the 6-h forecast ensemble variance (i.e., the ensemble spread) at locations of the rawinsonde observations.

a 95% confidence interval, were plotted. For both the temperature and wind speed, MR170 has a closer fit to the observations when compared with SR-High for all vertical levels (Figs. 6a,b). The improvement is most apparent in the upper levels of the TR region for both the temperature and wind speed. Although there is a marginal improvement for specific humidity fit to observations globally, the improvement is not statistically significant (Fig. 6c), suggesting the larger localization length scales and greater sampling at larger scales in MR170 does not provide substantial benefit to specific humidity as for temperature and wind speed.

Additionally, the spread of the ensemble, ideally, should be representative of the ensemble forecast errors. However, ensemble-based DA systems often suffer from underdispersiveness of the background ensemble due to the deficiency of the DA system and the misrepresentation of the model errors (e.g., Houtekamer and Mitchell 1998; Whitaker and Hamill 2002). In general, the total RMSF of the 6-h forecast to observations in Fig. 7 shows that the temperature errors are largest at the surface and upper levels for all regions, while wind speed and specific humidity errors tend to increase with decreasing pressure. Consistent with Fig. 6, the MR170 experiment slightly outperforms the SR-High experiment for all regions and levels. When comparing the RMSF to the total spread (i.e., the square root of the sum of the 6-h background ensemble variance and the observation error variance), the ensemble is shown to be underdispersive for all variables. However, this difference between the RMSF and spread is smaller for the MR170 experiment than for the SR-High experiment.

Figure 8 further demonstrates how the ensemble spread is increased in the MR170 experiment for most vertical levels for temperature and wind speed and lower levels for specific humidity. This increase is most apparent in the SH, especially in levels that are typically associated with substantial temporal and spatial variance for each variable (i.e., near the surface for temperature and specific humidity and near jet-level for wind speed). This increase may be counterintuitive since ensemble-based DA methods using higher resolutions, as is used by all ensemble members in the SR-High experiment in this study, typically lead to increased ensemble spread (e.g., Pellerin et al. 2003; Hamrud et al. 2015). However, the increase in spread for the temperature, wind speed, and low-level specific humidity for the MR170 experiment, in part, might be due to the increase in the total number of ensemble members. Increased ensemble size typically leads to a decrease in sampling error, which could lead to a spread more representative of the error of the system—corresponding to a larger ensemble spread in an underdispersive ensemble. The difference in spread is also more prevalent in dynamically or thermodynamically active regions with sharp gradients, or greater uncertainty. For example, over the experiment period, the regions associated with the greatest increase in spread at 200 hPa consistently correspond with the strongest jet regions (not shown). Compared to wind and temperature, specific humidity tends to have larger spread using higher resolution ensembles (Fig. 8c) at upper levels, likely due to its relatively small-scale nature.

c. *Forecast errors at longer lead times*

To further evaluate SR-High versus MR170, errors of the deterministic forecasts initialized by the analysis produced by these two approaches are calculated. The RMSF of the forecasts at 1° grid spacing using bilinear interpolation to ERA-Interim (1° × 1°; Berrisford et al. 2011) was calculated at each forecast lead time. Since the magnitude of forecast errors and errors in reanalyses can be comparable at early lead times (Bormann et al. 2019), verification using ERA-Interim will not be shown prior to a 24-h lead time (e.g., Magnusson and Källén 2013; Arcomano et al. 2020; Chen 2020; Kay and Wang 2020). The temporally averaged difference in the RMSF to the reanalysis for temperature and wind speed between the MR170 and the SR-High experiments for 1–5-day lead times in 6-h increments and 16 vertical pressure levels is shown for the GB, NH, SH, and TR regions in Fig. 9 with statistical significance determined by a paired t test with a 95% confidence interval. Globally, MR170 demonstrates improvement for nearly all vertical levels and lead times for both temperature and wind speed (Figs. 9a,b). The magnitude of this improvement generally increases for later lead times with the greatest improvement predominately near jet level. The

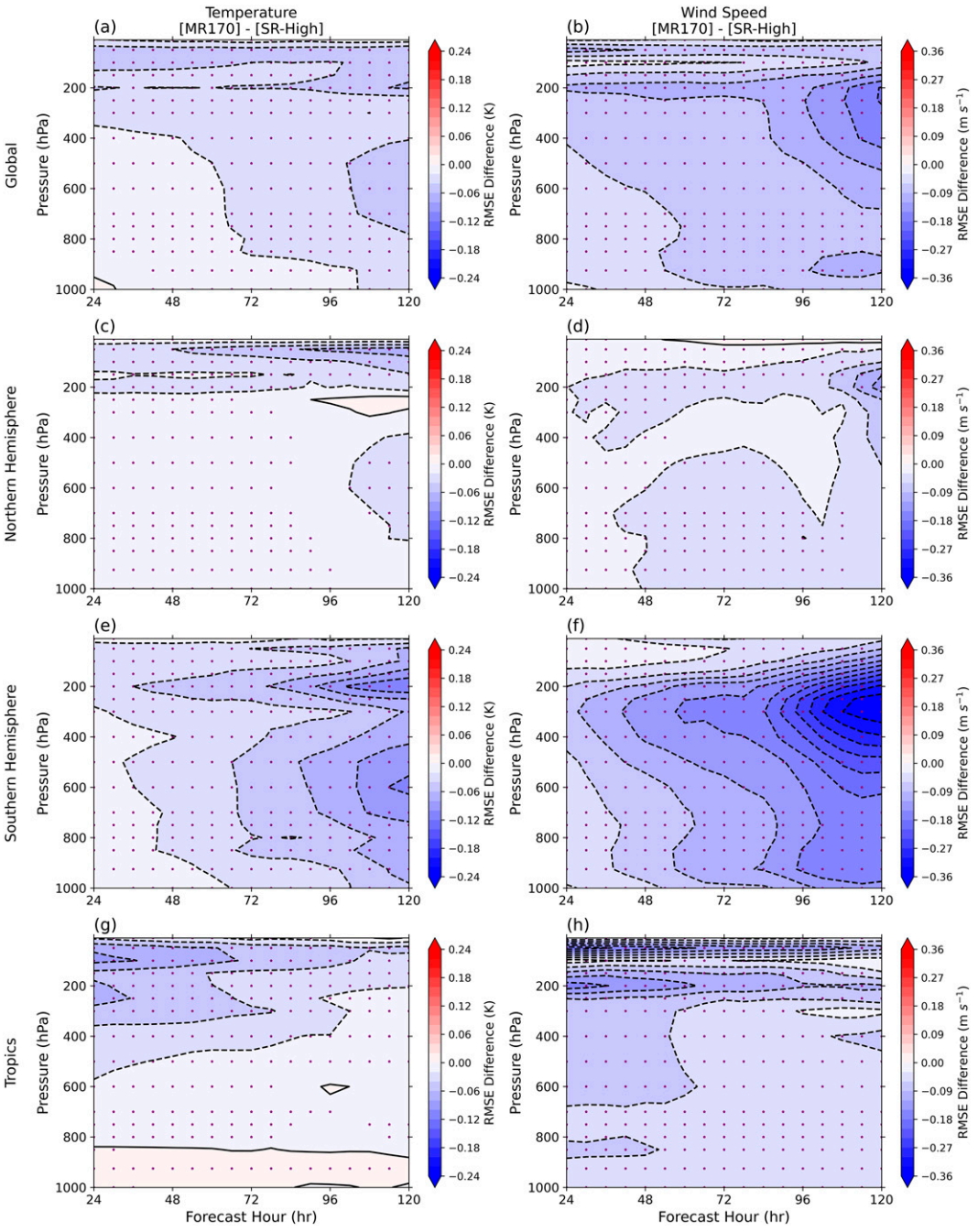

FIG. 9. Difference in RMSF of the MR170 and SR-High forecasts to the ERA-Interim (left) temperatures (K) and (right) wind speeds (m s^{-1}) averaged over all cycles for (a),(b) global; (c),(d) Northern Hemisphere extratropical; (e),(f) Southern Hemisphere extratropical; and (g),(h) tropical latitudes for 1–5 days' lead time. Red and blue shading indicate a larger RMSF to the reanalysis for MR170 and SR-High, respectively. Purple dots indicate levels where the RMSF difference is statistically significant using a paired *t* test and a 95% confidence interval.

differences in the RMSF to the reanalysis for the NH and SH regions mainly resemble the global pattern (Figs. 9c–f). The magnitude of the improvement in the SH, however, is much larger than in the NH for the same lead time and level. MR170 also has general improvement in the TR region with the exception of the temperature at the lowest levels (Figs. 9g,h). However, the largest improvement in the TR region is at the early lead times rather than later lead times as in other regions.

To determine the multiscale impact of the DA methodologies on the forecast errors, the temperature and wind errors for

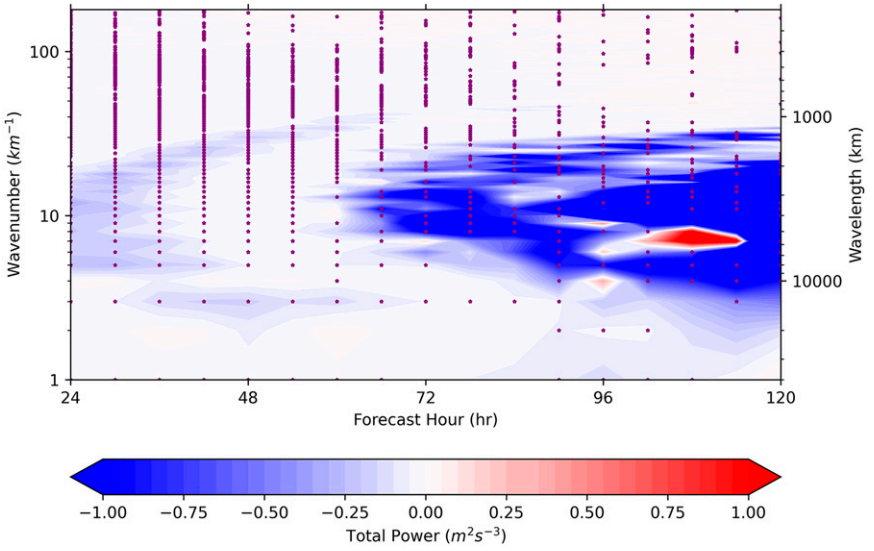

FIG. 10. Difference in total power error (m^2 s^{-3}) between the MR170 and SR-High experiments verified against ERA-Interim, averaged over all DA cycles and pressure levels selected in Fig. 9 and given in terms of wavenumber (left y axis) and corresponding wavelength (right y axis) for 1–5 days' lead time. Red and blue shading indicate larger total power errors for MR170 and SR-High, respectively. Purple dots indicate levels where the difference is statistically significant using a paired t test and a 95% confidence interval.

each vertical level and lead time shown in Fig. 9 were decomposed using spherical harmonics. The total energy norm (e.g., Wang and Bishop 2003) was then calculated with the formula:

$$\frac{1}{2}u'^2 + \frac{1}{2}v'^2 + \frac{c_p}{T_r}T'^2, \tag{3}$$

where u', v', and T' are the zonal wind, meridional wind, and temperature errors, respectively, relative to ERA-Interim, and c_p and T_r are the specific heat at constant pressure at 1.0057 J kg^{-1} K^{-1} and the reference temperature at 270 K, respectively (Ehrendorfer et al. 1999).

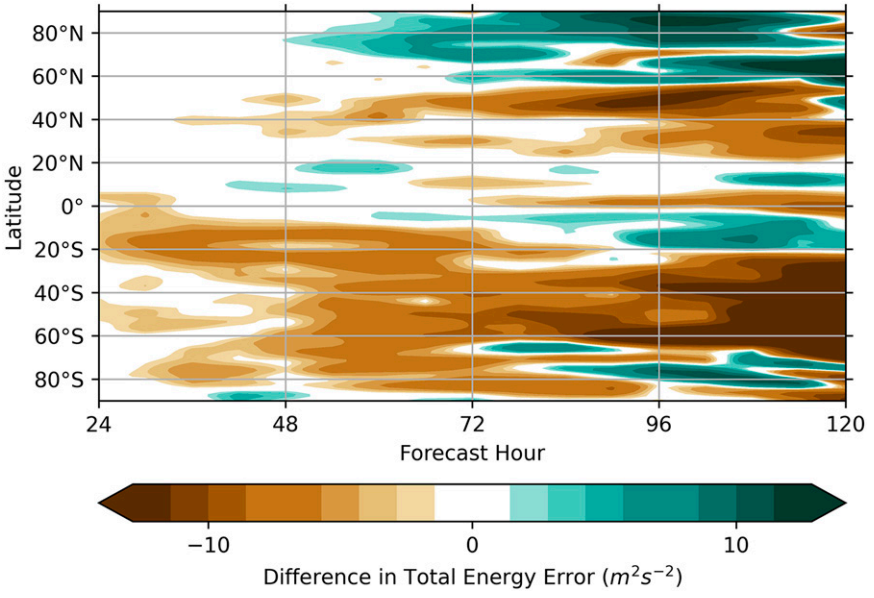

FIG. 11. Difference in total energy error (m^2 s^{-2}) at 250 hPa filtered to include wavenumbers 5–25 for the 0000 UTC 12 Sep 2017 cycle between the MR170 and SR-High experiments verified against ERA-Interim and averaged over all longitudes at a given latitude and 1–5 days' lead time.

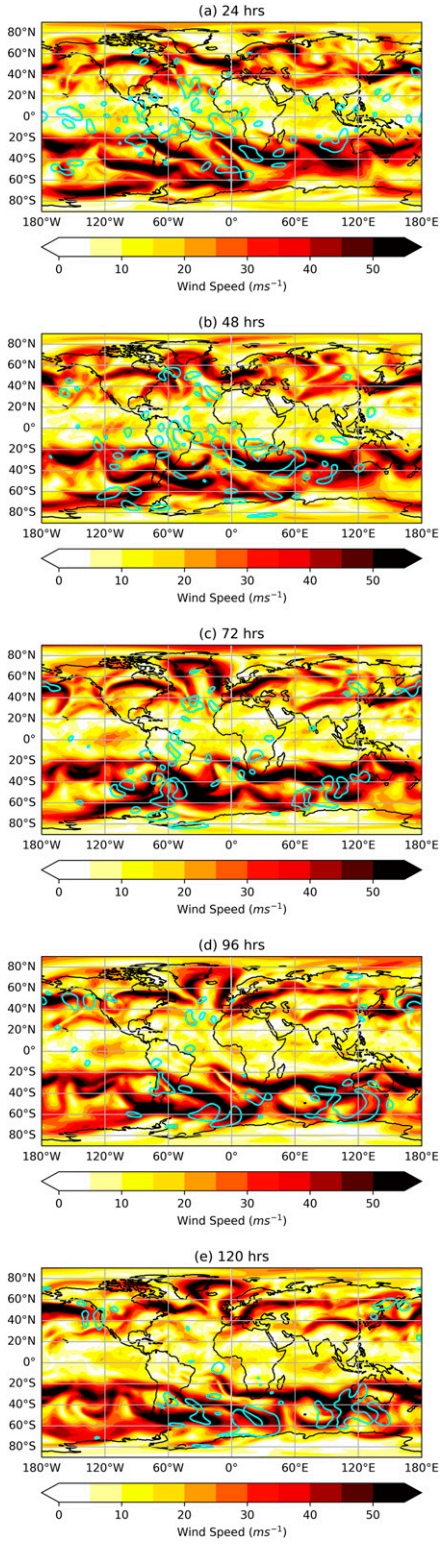

FIG. 12. ERA-Interim wind speed at 250 hPa for the 0000 UTC 12 Sep 2017 cycle for every 24-h forecast up to 5 days. The 5% minimum difference in total energy error at 250 hPa relative to ERA-Interim between the MR170 and SR-High experiments filtered to include wavenumbers 5–25 is contoured in cyan.

The difference between the spectrally decomposed errors for the two experiments averaged over the duration of the experiment is shown in Fig. 10. At 24 h, the MR170 experiment shows improvement in errors at nearly all wavenumbers with a maximum between wavenumbers 5 and 25 (wavelengths of ~8000–1600 km). This peak disperses into larger and smaller wavenumbers, and a second maximum occurs at similar wavenumbers starting right before 72 h. From 72 to 120 h, these differences in errors continue to grow as well as disperse into surrounding wavenumbers.

To diagnose the physical features that led to the differences in the forecast errors between the two experiments, one case, which was representative of the errors shown in Figs. 9 and 10, was chosen. The case had an analysis time of 0000 UTC 12 September 2017, and the 250-hPa level was chosen to be analyzed since pressure levels between 200 and 300 hPa typically demonstrated the greatest improvement in Fig. 9. To determine specifically what was occurring at the peaks in the forecast errors in Fig. 10, the total energy error was calculated at each grid point for each experiment and lead time. Then, a filter was applied to extract the wavenumbers between 5 and 25 since these wavenumbers displayed the greatest improvement in Fig. 10. In Fig. 11, these results were averaged over all longitudes, and the difference between MR170 and SR-High was taken. Figure 12 displays the locations of the 5% greatest improvement for the MR170 experiment, filtered for the same wavenumbers as well as the associated 250-hPa wind speeds in 24-h increments.

At 24 h, there is peak in the tropics, especially near and south of the equator, of the difference in energy errors, indicating improvement for MR170 in this region (Fig. 11). One of the largest areas of improvement, spanning from just north of the equator to around 20°S and from about 20°W to 5°E, encompasses the western portion of the tropical easterly jet (TEJ; Koteswaram 1958) and extends southward to the northern portion of the subtropical jet (Fig. 12a). Improvement at 24 h is likely due to large-scale improvement at analysis time within or upstream of these features. Between 24 and 72 h, locally generated errors in the NH and SH extratropics begin to become evident (Figs. 11, 12a–c). These errors are largely collocated with regions of the polar jet in both hemispheres (e.g., around 60°S, 120°W at 48 h). Additionally, areas where the polar and subtropical jets interact in the SH tend to exhibit more substantial growth of the differences in errors, such as shown at 72 h near 50°S, 50°W or near 50°S, 105°E. After 72 h, the error differences continue to increase in both hemispheres but more so in the SH than the NH. This growth is especially apparent in polar jet region in the SH, especially in the areas to southwest of Australia (~40°S, 110°E), south of Africa (~60°S, 35°E), and southeast of South America (~50°S, 30°W; Figs. 12d,e).

The case study highlights the value in accurately analyzing large-scale features such as the TEJ as well as the importance of the interaction between the tropics, subtropics, and extratropics in terms of error growth. For this case, the NH jet stream is relatively weak and isolated from the tropical region as would be expected in the late summer and early fall.

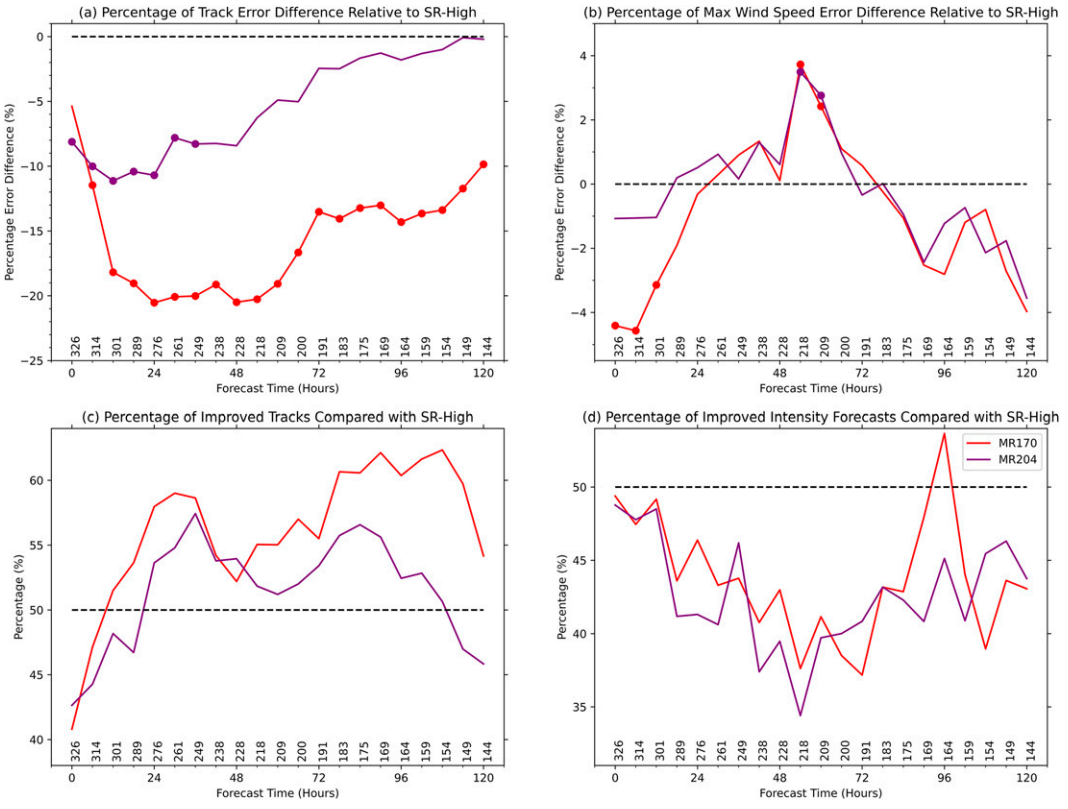

FIG. 13. (top) Percentage of (a) track and (b) intensity forecast error differences of MR170 (red) and MR204 (purple) relative to SR-High for lead times up to 5 days. Dots indicate statistically significant differences using a paired *t* test and a 95% confidence interval. (bottom) Percentage of more accurate MR-ENS (c) track and (d) intensity forecasts compared with SR-High using the same colors and lead times as in (a) and (b). The numbers above the *x* axes indicate the sample size of tracks at the corresponding lead time.

Therefore, error growth in this region depends mainly on errors that are generated locally. In contrast, the SH subtropical jet directly interacts with regions with large differences in errors in the tropics, which leads to subsequent error growth along the subtropical jet. In areas where the subtropical and polar jet meet, the locally generated errors in the SH extratropics are further compounded by the errors at early lead times in the tropics and subtropics.

The large-scale improvements in the forecasts for MR170 relative to SR-High can be attributed to two interconnected factors: 1) the larger number of low-resolution ensemble members in MR170; 2) the larger localization radius for the low-resolution ensemble. Both factors can lead to a more accurate estimation of the background error covariances at relatively large scales and therefore more accurate correction for the background forecast for the large scales.

d. Tropical cyclone track errors

In addition to examining global forecasts, the impact of the extended MR-ENS methodology on TC track and intensity forecast errors was also investigated. For each 5-day forecast, a tracking algorithm, as described in Marchok (2002), was used to determine the forecasted location and maximum wind speeds of 15 TC storms that occurred during the experiment period. These storms include Irma, Jose, Katia, Lee, and Maria in the Atlantic basin; Lidia, Max, Norma, Otis, and Pilar in the east Pacific basin; and Sanvu, Mawar, Guchol, Talim, and Doksuri in the west Pacific basin. TC track errors were calculated by taking the difference of the forecasted track or maximum wind speeds and the "best track" data as determined by the NHC[2] for the Atlantic and east Pacific basins and the JTWC[3] for the west Pacific basin. Track and intensity errors were averaged over all 15 storms for each experiment and lead time. The statistical significance for differences between MR-ENS and SR-High track and intensity errors was calculated by using a paired *t* test with a 95% confidence interval.

In Fig. 13a, the percentage difference in temporally averaged track errors for MR-ENS experiments with SR-High used as the reference is shown. Between 6 and 120 h, the MR170 experiment has a statistically significant decrease in TC track errors. The largest improvements occur between 12 and 54 h where the errors decrease by ~20%. The

[2] https://www.nhc.noaa.gov/data/#hurdat.
[3] https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks.

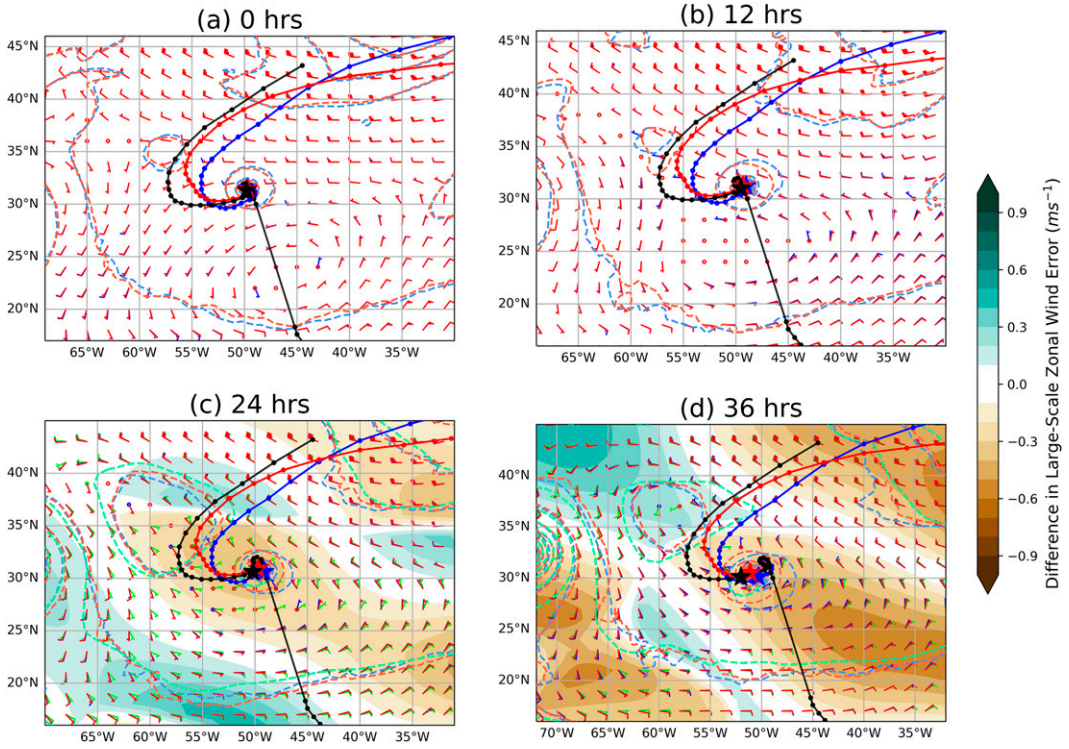

FIG. 14. Forecasted tracks of Tropical Cyclone Lee for MR170 (red solid) and SR-High (blue solid) for lead times of (a) 0, (b) 12, (c) 24, and (d) 36 h with an initialization time of 1800 UTC 24 Sep 2017, along with the best track of the storm (black solid). Dots are in 6-h increments for the tracks. Stars indicate the central position of the storm at the indicated lead times for the MR170 (red), SR-High (blue), and best (black) tracks. Mean sea level pressure is plotted every 4 hPa for the MR170 (red dashed), SR-High (blue dashed), and ERA-Interim (green dashed at 24 and 36 h). Wind barbs, indicating the large-scale (wavenumbers 0–10) pressure-weighted 850–300-hPa wind speed (kt; 1 kt $\approx$ 0.51 m s^{-1}) and direction, are plotted for MR170 (red), SR-High (blue), and ERA-Interim (green at 24 and 36 h) at each lead time. At 24 and 36 h, shading indicates the difference in the magnitude of the pressure-weighted 850–300-hPa zonal wind error filtered to include wavenumbers 0–10 between MR170 and SR-High relative to ERA-Interim, where green and brown indicate larger errors for MR170 and SR-High, respectively.

percentage of tracks that have smaller errors compared with SR-High are shown in Fig. 13c, where a 100% decrease would indicate that the MR-ENS experiment has smaller track errors at a certain lead time for all TC tracks, while 50% would signify an equal likelihood of MR-ENS or SR-High having smaller track errors if a track at a given lead time were randomly selected. For MR170, the percentage of track errors that are decreased compared with SR-High exceeds 50% for lead times between 12 and 120 h.

In terms of intensity, the experiments have mixed improvement in error magnitude with significant improvement for MR170 at 0–12 h and for SR-High from 48 to 54 h (Fig. 13b). These improvements are relatively small at less than 5% for all lead times. MR170 also improves the intensity forecast less than 50% of the time for all forecast hours except 96 h (Fig. 13d). The mixed performance of MR170 and SR-High for intensity relative to track is consistent with hurricane intensity having a stronger dependence on smaller-scale processes.

In the 0–6-h timeframe, however, there are fewer total tracks that are improved both at and immediately after analysis time for the MR170 experiment (Fig. 13c). Also, the percentage difference between track errors are larger at 12–54-h lead times when compared to 0–6-h lead times (Fig. 13a). As discussed in section 4a, the SR-High experiment has larger kinetic energy analysis increment power, especially at medium- to smaller-scales in the Atlantic (Fig. 4). Though the track forecast of a TC is mainly dominated by larger-scale flow, smaller-scale features may still affect the analysis position of the storm and thereby subsequent short-term forecasts. The larger corrections at small scales for the SR-High experiment may help explain the relative degradation of the MR170 track forecasts in the 0–6-h timeframe as well as why SR-High has more tracks that are improved during this time. At very short lead times, there is a balance between resolving and correcting smaller-scale features—as may be accomplished with SR-High—as well as correcting larger-scale features with greater accuracy—for which MR170 is better suited. At later lead times, the large-scale correction associated with MR170 dominates the improvement in the track forecasts.

To determine the cause of the differences in track errors, a single track, which showed improvements exceeding 50%

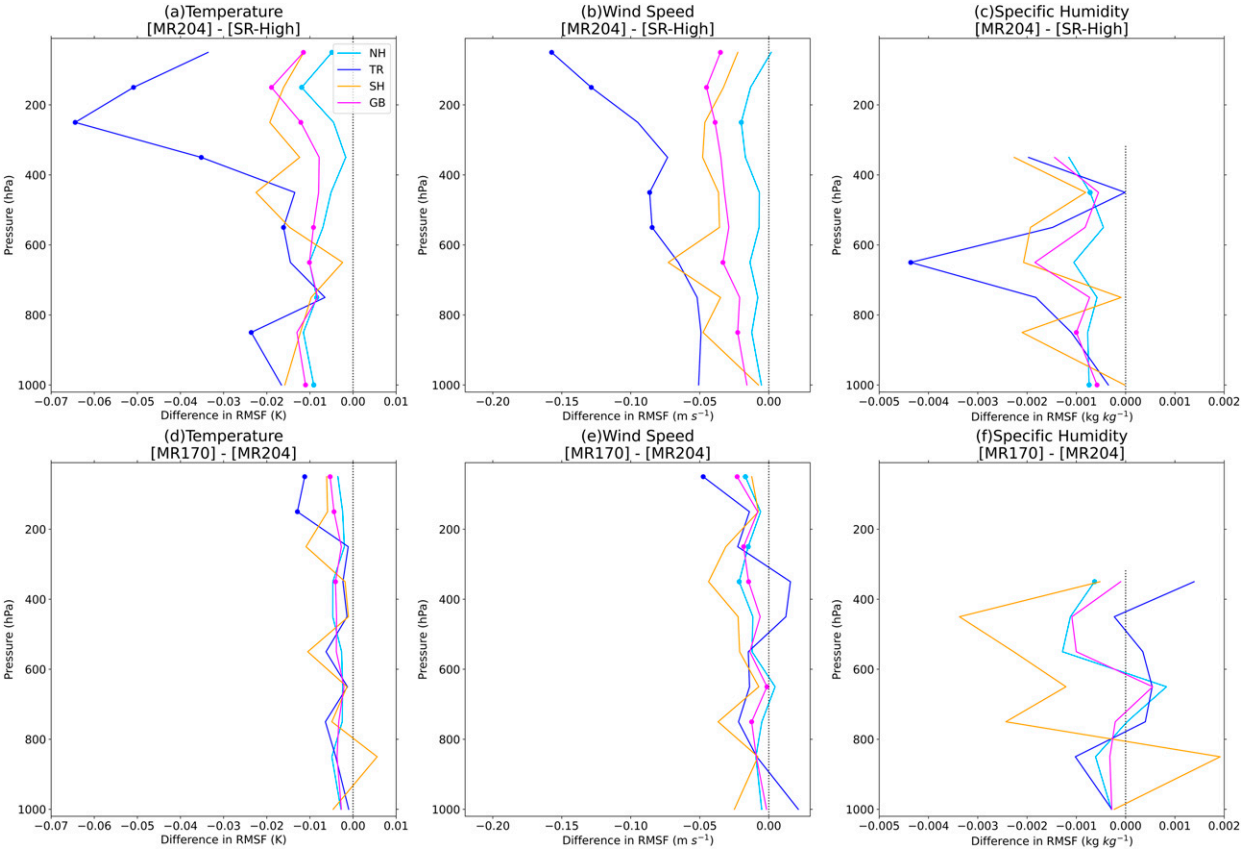

FIG. 15. As in Fig. 6, but (a)–(c) between the MR204 and SR-High experiments and (d)–(f) between the MR170 and MR204 experiments.

in the 0–120-h track forecast for the MR170 experiment, was selected for further examination. This track was associated with Hurricane Lee with an analysis time of 1800 UTC 24 September 2017 soon after its re-intensification into hurricane strength (Blake 2018).

The track of a tropical cyclone is largely determined by the steering flow, which can be estimated by the environmental wind field averaged vertically over several pressure levels (e.g., George and Gray 1976; Chan and Gray 1982). For this case, a pressure-weighted average of the zonal and meridional winds at each grid point from 850 to 300 hPa (e.g., Wu and Chen 2016) was calculated. The effects of the storm itself then were removed by filtering the winds to include only large scales (approximately wavenumbers 0–10). These large-scale steering flow winds are plotted with the track positions in Fig. 14. The difference in the magnitude of the error of the large-scale zonal winds between the MR170 and SR-High experiments relative to ERA-Interim are also plotted for lead times beginning at 24 h.

Compared with SR-High, MR170 improves the analysis of the central position of the storm, which is located at the northern portion of a large-scale anticyclonic region (Fig. 14a). MR170 produces weaker westerly winds in the analysis in these regions, leading to the central location of Lee in the MR170 12-h forecast remaining further to the east, closer to the location of the best track (Fig. 14b). During subsequent lead times, the large-scale anticyclone propagates northwestward, shifting the direction of the large-scale winds in the vicinity of the storm center from westerly to easterly (Figs. 14b–d). From 24 to 36 h, the large-scale zonal wind errors associated with the anticyclonic region are smaller for MR170, likely stemming from an improved analysis of these large-scale winds. The smaller errors correspond with stronger easterly large-scale winds and larger westward storm motion from 18 to 36 h, more consistent with the best track. These improvements can be connected to the ability of MR170 to analyze more accurately the large-scale wind patterns due to the increased number of low-resolution ensemble members and the larger localization radius associated with the low-resolution ensemble.

e. *Sensitivity testing*

To further test the sensitivity of MR-ENS to the variable sizes of the high- and low-resolution ensemble members, a second MR-ENS experiment with 180 low-resolution members and 24 high-resolution members (MR204) was examined. As discussed in section 3, MR204 has the same cost as MR170 and SR-High.

The MR204 500-hPa analysis increments decomposed into spectral space show a similar pattern to that of MR170 for all variables (Fig. 3). Both MR170 and MR204 show a

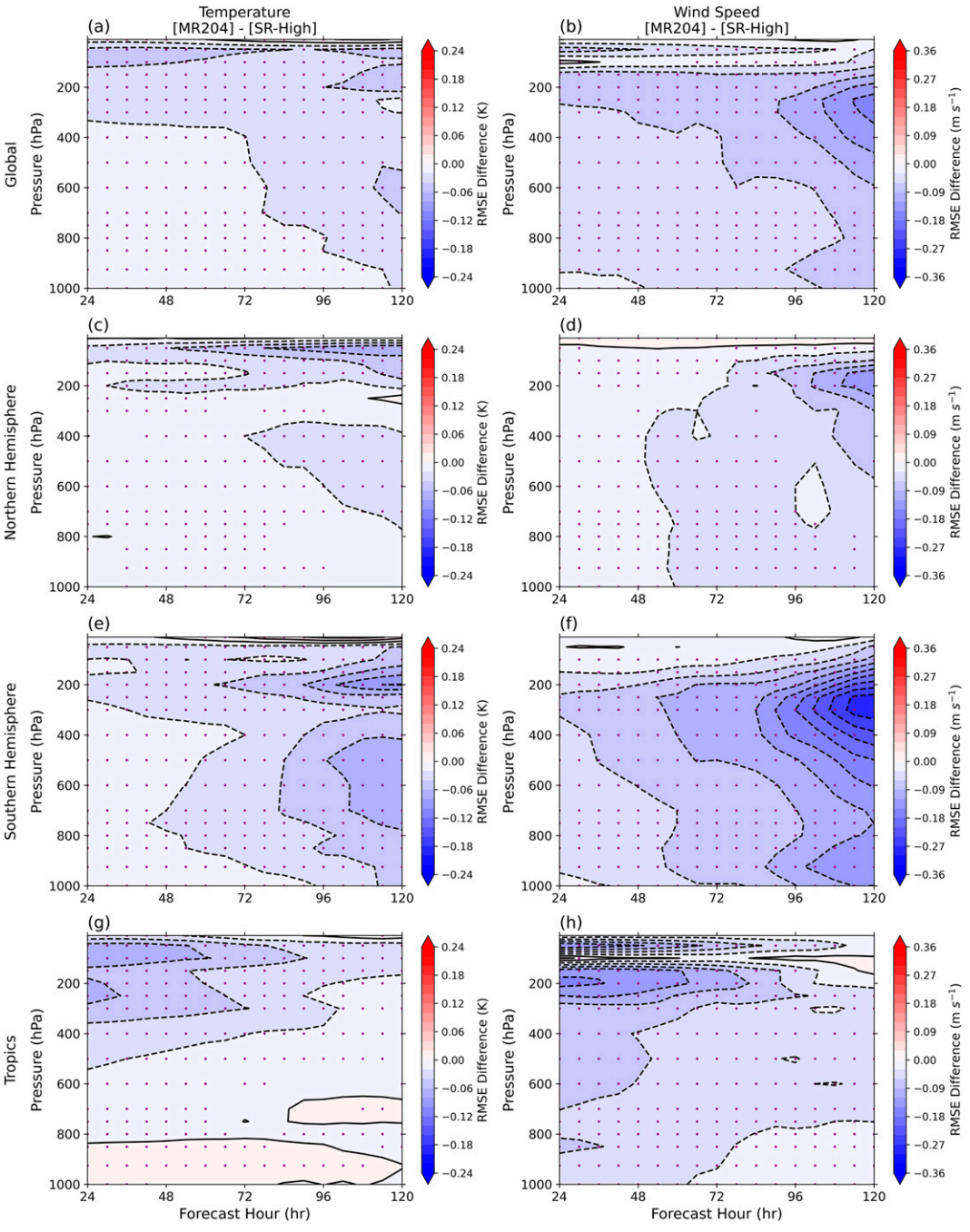

FIG. 16. As in Fig. 9, but for the MR204 and SR-High experiments.

substantial increase in the analysis increments over SR-High at larger spatial scales for temperature and zonal wind and similar analysis increments as SR-High for specific humidity. All variables have similar or slightly smaller analysis increments at smaller scales.

Like MR170, MR204 produced an improved RMSF of the 6-h global forecast to the rawinsonde observations, compared with SR-High for the temperature, wind speed, and specific humidity at all vertical levels (Figs. 15a–c). MR204, however, slightly degrades the accuracy of the 6-h forecast when compared with MR170 for most vertical levels (Figs. 15d–f). Similarly, at longer lead times, MR204 improves the RMSF to the reanalysis compared to SR-High (Fig. 16). However, MR170 generally exhibits significant improvement in the global forecast over MR204 at most vertical levels and lead times (Fig. 17). For TC track forecasts, MR204 produces smaller track errors compared with SR-High up to 120 h with statistically significant differences up to 36 h (Fig. 13a). Also,

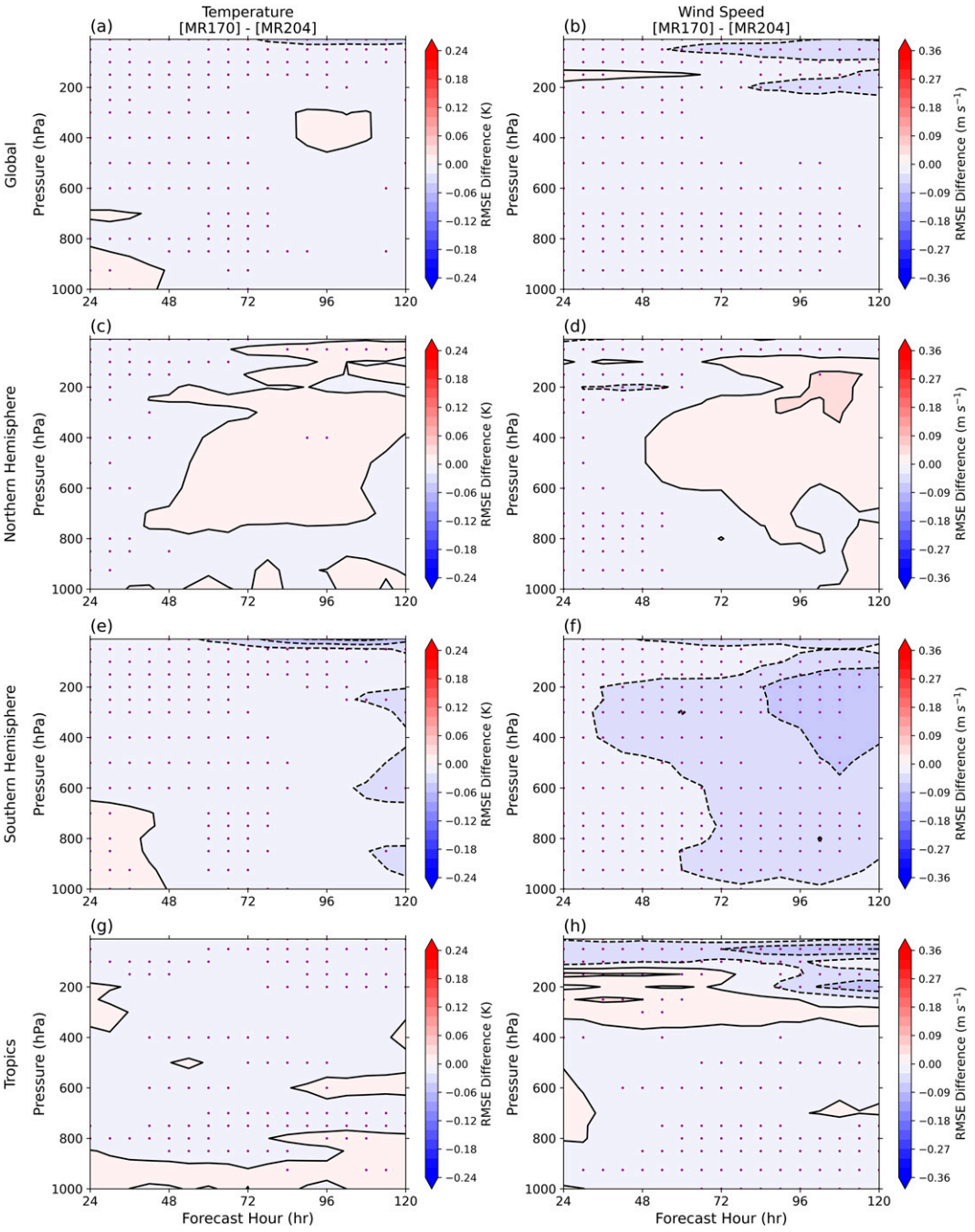

FIG. 17. As in Fig. 9, but for the MR170 and MR204 experiments.

between 24 and 108 h MR204 improves the track forecast more than 50% of the time when compared with SR-High (Fig. 13c). For nearly all lead times, however, there is a degradation in the track forecast for MR204 when compared to MR170. Compared with SR-High, TC intensity errors follow a similar pattern for MR170 and MR204. MR170 shows similar values or improvement over MR204 at most lead times (Figs. 13b,d).

Overall, both MR-ENS experiments improve upon SR-High. However, MR204 shows degradation when compared with MR170. Compared to MR170, MR204 has increased low-resolution ensemble members and decreased high-resolution members. The degraded performance of MR204 is likely due to the loss of information at smaller spatial scales outweighing the additional gain of resolving large-scale errors. The need to include certain numbers of high-resolution

ensemble members in the background error sampling is a result consistent with Lei and Whitaker (2017). These sensitivity experiment results demonstrate that optimal performance of MR-ENS requires a balanced selection of high- and low-resolution members.

5. Conclusions

This study further develops the multiresolution ensemble (MR-ENS) hybrid 4DEnVar. The extended MR-ENS method allows for a varying number of low- and high-resolution ensemble members as well as different localization length scales for the low- and high-resolution ensembles.

The extended MR-ENS technique was evaluated with 6-h cycling for a 5-week period. Two MR-ENS experiments were examined. The first, MR170, had 130 low-resolution members and 40 high-resolution members; and the second, MR204, had 180 low-resolution members and 24 high-resolution members. The deterministic analysis and forecast for these experiments were compared to those from a single high-resolution ensemble hybrid 4DEnVar experiment, SR-High, at the same computational cost.

Globally, larger analysis increment power was present for MR-ENS experiments at large scales, especially for temperature and zonal wind. At the smallest scales, MR-ENS experiments exhibited slightly smaller increment power. In medium scales, MR-ENS produced larger analysis increments in most regions, most notably the SH, where large-scale forcing had the largest impact on the region during the experiment period. There was an exception to this notion in the tropical and NH Atlantic, which were dominated by high-intensity convective activity during this period with contributions from five hurricanes (including four major hurricanes) and a substantial number of tropical easterly waves.

The MR-ENS experiments generally led to an improvement in the deterministic forecast for up to five days in lead time when compared to SR-High, despite the experiments having the same computational cost. This improvement was most apparent in large scales and upper levels in the tropics at early lead times (prior to 48 h) and the SH at later lead times. A case study highlighted that the greatest improvement occurred after the interaction of large-scale features, such as the interaction of the TEJ with the subtropical jet or the interaction of the subtropical and polar jets. Therefore, it is beneficial to create an analysis that can both correct the large-scale features themselves and allow for more multiscale interaction between these features.

As with global forecasts, tropical cyclone track forecasts were generally improved for the MR-ENS experiments, with improvements exceeding 10% on average for most lead times. The improvements were mainly attributed to the ability of the extended MR-ENS technique to further correct the large-scale environmental flow due to greater sampling at large scales and broader localization radius for the low-resolution ensemble. At very early lead times (0–6 h), there was less notable improvement in track forecasts when compared with later lead times, and a greater number of total tracks were improved by the SR-High experiments over MR-ENS. The slight degradation in MR-ENS from 0 to 6 h compared to other lead times corresponds with the larger analysis increments for SR-High in TC basins during the experiment period. Additionally, for nearly all lead times, there were fewer intensity forecasts improved by MR-ENS than SR-High, different from the track forecasts. Therefore, there are still applications for which having greater information at high-resolution via more high-resolution ensemble members is preferable to having greater sampling at low resolution.

The two MR-ENS experiments were intercompared. The MR-ENS configuration with a smaller (larger) number of low- (high-) resolution ensemble members, MR170, outperformed its counterpart, MR204, in most metrics. This result shows that it is not enough to rely solely on the corrections at large scales. There is still evident forecast value in retaining more accurate information about the state and error correlations at small scales by having a sufficient sample of high-resolution ensemble members. The optimum balance between correcting large and small scales at a given computational cost can be found by adjusting the size of the low- and high-resolution ensembles using a MR-ENS framework. The exact values associated with that optimal balance are beyond the scope of this study. However, the optimal number of high- and low-resolution members seems to be feature-dependent. Therefore, it would be useful to further test the extended MR-ENS capabilities at additional times and under different conditions outside of the 5-week period used in these experiments (e.g., during a NH winter). This study is the first to compare SR-High and MR-ENS at the same costs. Tuning of additional MR-ENS parameters may be needed to optimize its performance. For example, tuning the hybrid weights for the low- and high-resolution ensembles or formulating the weights to be scale-dependent may be beneficial to more accurately correct different scales. This study, also, only tested two different horizontal resolutions in a deterministic context. Further extension of MR-ENS to include three or more resolutions and to include variable vertical resolution could allow an even greater flexibility of sampling of various scales.

This study focuses only on MR-ENS. Combining MR-ENS with additional multiscale DA techniques, such as SDL or other methods (e.g., Buehner and Shlyaeva 2015; Wang et al. 2021), requests additional algorithm development. The relative difference between MR-ENS and SR-High when SDL is implemented and integrated should be examined in future work. Last, cautions need to be taken to extrapolate the results to the operational system given the overall coarser resolutions used in this study.

Acknowledgments. This study was supported by NOAA Awards NA18NWS4680051 and NA22NWS4680011. Additional funding for the first author was provided by an American Meteorological Society (AMS) Graduate Fellowship. Special thanks go to Yongming Wang for assistance with GSI and the FV3GFS. Some results and descriptions were included in the first author's master's thesis and the abstract of the authors' conference presentations. All computational resources for this

study were provided by NOAA's Research and Development High-Performance Computing System Jet machine.

Data availability statement. The data for the experiments performed in this study are archived locally and available upon request to the corresponding author.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

Arcomano, T., I. Szunyogh, J. Pathak, A. Wikner, B. R. Hunt, and E. Ott, 2020: A machine learning-based global atmospheric forecast model. *Geophys. Res. Lett.*, **47**, e2020GL087776, https://doi.org/10.1029/2020GL087776.

Balle, T., and P. Johnsen, 2016: Improving I/O performance of the Weather Research and Forecast (WRF) Model. Scalability CUG 2016, ECMWF, 9 pp., https://cug.org/proceedings/cug2016_proceedings/includes/files/pap123s2-file1.pdf.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Benjamin, S. G., J. M. Brown, G. Brunet, P. Lynch, K. Saito, and T. W. Schlatter, 2019: 100 years of progress in forecasting and NWP applications. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial, Meteor. Monogr.*, No. 59, Amer. Meteor. Soc., https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1.

Berrisford, P., and Coauthors, 2011: The ERA-Interim archive version 2.0. ECMWF, 27 pp., https://www.ecmwf.int/node/8174.

Bishop, C. H., 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.

Blake, E. S., 2018: National Hurricane Center tropical cyclone report: Hurricane Lee (14-30 September 2017). NHC Tech Rep. AL142017, NHC, 16 pp., https://www.nhc.noaa.gov/data/tcr/AL142017_Lee.pdf.

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.

Bormann, N., H. Lawrence, and J. Farnan, 2019: Global observing system experiments in the ECMWF assimilation system. ECMWF Tech. Memo. 839, 26 pp., https://www.ecmwf.int/en/elibrary/80953-global-observing-system-experiments-ecmwf-assimilation-system.

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.

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

Burpee, R. W., 1972: The origin and structure of easterly waves in the lower troposphere of North Africa. *J. Atmos. Sci.*, **29**, 77–90, https://doi.org/10.1175/1520-0469(1972)029<0077:TOASOE>2.0.CO;2.

Carlson, T. N., 1969: Some remarks on African disturbances and their progress over the tropical Atlantic. *Mon. Wea. Rev.*, **97**, 716–726, https://doi.org/10.1175/1520-0493(1969)097<0716:SROADA>2.3.CO;2.

Chan, J. C. L., and W. M. Gray, 1982: Tropical cyclone movement and surrounding flow relationships. *Mon. Wea. Rev.*, **110**, 1354–1374, https://doi.org/10.1175/1520-0493(1982)110<1354:TCMASF>2.0.CO;2.

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.

Chen, Y., 2020: COVID-19 pandemic imperils weather forecast. *Geophys. Res. Lett.*, **47**, e2020GL088613, https://doi.org/10.1029/2020GL088613.

Domingues, M. O., O. Mendes, and A. M. da Costa, 2005: On wavelet techniques in atmospheric sciences. *Adv. Space Res.*, **35**, 831–842, https://doi.org/10.1016/j.asr.2005.02.097.

Ehrendorfer, M., R. M. Errico, and K. D. Raeder, 1999: Singular-vector perturbation growth in a primitive equation model with moist physics. *J. Atmos. Sci.*, **56**, 1627–1648, https://doi.org/10.1175/1520-0469(1999)056<1627:SVPGIA>2.0.CO;2.

Engel, T., A. H. Fink, P. Knippertz, G. Pante, and J. Bliefernicht, 2017: Extreme precipitation in the West African cities of Dakar and Ouagadougou: Atmospheric dynamics and implications for flood risk assessments. *J. Hydrometeor.*, **18**, 2937–2957, https://doi.org/10.1175/JHM-D-16-0218.1.

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.

George, J. E., and W. M. Gray, 1976: Tropical cyclone motion and surrounding parameter relationships. *J. Appl. Meteor. Climatol.*, **15**, 1252–1264, https://doi.org/10.1175/1520-0450(1976)015<1252:TCMASP>2.0.CO;2.

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.

——, and J. S. Whitaker, 2005: Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Mon. Wea. Rev.*, **133**, 3132–3147, https://doi.org/10.1175/MWR3020.1.

——, ——, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

——, ——, M. Fiorino, S. E. Koch, and S. J. Lord, 2010: Increasing NOAA's computational capacity to improve global forecast modeling: A NOAA white paper, 25 pp., https://www.psl.noaa.gov/people/tom.hamill/global_HPC_whitepaper_hamilletal.pdf.

Hamrud, M., M. Bonavita, and L. Isaksen, 2015: EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. *Mon. Wea. Rev.*, **143**, 4847–4864, https://doi.org/10.1175/MWR-D-14-00333.1.

Harris, L., X. Chen, W. M. Putman, L. Zhou, and J.-H. Chen, 2021: A scientific description of the GFDL finite-volume cubed-sphere dynamical core. NOAA Tech. Memo. OAR GFDL, 2021-001, 109 pp., https://doi.org/10.25923/6nhs-5897.

Hayden, C. M., and R. J. Purser, 1995: Recursive filter objective analysis of meteorological fields: Applications to NESDIS

operational processing. *J. Appl. Meteor. Climatol.*, **34**, 3–15, https://doi.org/10.1175/1520-0450-34.1.3.

Hollis, M., 2021: Global analysis of characteristics of tropical easterly waves and related precipitation. M.S. thesis, School of Meteorology, University of Oklahoma, 87 pp., https://shareok.org/handle/11244/331416.

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.

——, X. Deng, H. L. Mitchell, S.-J. Baek, and N. Gagnon, 2014: Higher resolution in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 1143–1162, https://doi.org/10.1175/MWR-D-13-00138.1.

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.

——, ——, D. T. Kleist, and T. Lei, 2021: A simultaneous multiscale data assimilation using scale-dependent localization in GSI-based hybrid 4DEnVar for NCEP FV3-based GFS. *Mon. Wea. Rev.*, **149**, 479–501, https://doi.org/10.1175/MWR-D-20-0166.1.

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 Quart.*, No. 61, Fall 2018, 34 pp., https://doi.org/10.25923/JW00-R987.

Kay, J., and X. Wang, 2020: A multiresolution ensemble hybrid 4DEnVar for global numerical prediction. *Mon. Wea. Rev.*, **148**, 825–847, https://doi.org/10.1175/MWR-D-19-0002.1.

Kleist, D. T., and K. Ide, 2015: 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.

Koteswaram, P., 1958: The easterly jet stream in the tropics. *Tellus*, **10**, 43–57, https://doi.org/10.3402/tellusa.v10i1.9220.

Lei, L., and J. S. Whitaker, 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.

——, ——, and C. Bishop, 2018: Improving assimilation of radiance observations by implementing model space localization in an ensemble Kalman filter. *J. Adv. Model. Earth Syst.*, **10**, 3221–3232, https://doi.org/10.1029/2018MS001468.

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.

——, 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.

Lorenz, E. N., 1996: Predictability: A problem partly solved. Vol. 1, *Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1–19, https://www.ecmwf.int/node/10829.

Magnusson, L., and E. Källén, 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.*, **141**, 3142–3153, https://doi.org/10.1175/MWR-D-12-00318.1.

——, J.-H. Chen, S.-J. Lin, L. Zhou, and X. Chen, 2019: Dependence on initial conditions versus model formulations for medium-range forecast error variations. *Quart. J. Roy. Meteor. Soc.*, **145**, 2085–2100, https://doi.org/10.1002/qj.3545.

Marchok, T. P., 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.

Miyoshi, T., K. Kondo, 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.

Pellerin, G., L. Lefaivre, P. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes Geophys.*, **10**, 463–468, https://doi.org/10.5194/npg-10-463-2003.

Rainwater, S., and B. Hunt, 2013: Mixed-resolution ensemble data assimilation. *Mon. Wea. Rev.*, **141**, 3007–3021, https://doi.org/10.1175/MWR-D-12-00234.1.

Reed, R. J., D. C. Norquist, and E. E. Recker, 1977: The structure and properties of African wave disturbances as observed during phase III of GATE. *Mon. Wea. Rev.*, **105**, 317–333, https://doi.org/10.1175/1520-0493(1977)105<0317:TSAPOA>2.0.CO;2.

Torrence, C., and G. P. Compo, 1998: A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.*, **79**, 61–78, https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Walpole, R. E., R. H. Myers, S. L. Myers, and K. Ye, 2011: *Probability & Statistics for Engineers & Scientists*. Pearson, 816 pp.

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

——, 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 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.

——, C. H. Bishop, and S. J. Julier, 2004: Which is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590–1605, https://doi.org/10.1175/1520-0493(2004)132<1590:WIBAEO>2.0.CO;2.

——, C. Snyder, and T. M. Hamill, 2007: 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: Observing 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.*, **149**, 605–622, https://doi.org/10.1175/MWR-D-20-0290.1.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

——, and ——, 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.

Wu, L., and X. Chen, 2016: Revisiting the steering principal of tropical cyclone motion in a numerical experiment. *Atmos. Chem. Phys.*, **16**, 14 925–14 936, https://doi.org/10.5194/acp-16-14925-2016.

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916, https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2.

Ying, Y., F. Zhang, and J. L. Anderson, 2018: On the selection of localization radius in ensemble filtering for multiscale quasigeostrophic dynamics. *Mon. Wea. Rev.*, **146**, 543–560, https://doi.org/10.1175/MWR-D-17-0336.1.

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.