UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A MULTI-RESOLUTION ENSEMBLE HYBRID 4DENVAR WITH VARIABLE ENSEMBLE SIZES TO IMPROVE GLOBAL AND TROPICAL CYCLONE TRACK NUMERICAL PREDICTION

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

ERIN ASHLEY JONES
Norman, Oklahoma
2022

A MULTI-RESOLUTION ENSEMBLE HYBRID 4DENVAR WITH VARIABLE ENSEMBLE SIZES TO IMPROVE GLOBAL AND TROPICAL CYCLONE TRACK NUMERICAL PREDICTION

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Xuguang Wang, Chair

Dr. Steven M. Cavallo

Dr. James H. Rupert Jr.

Acknowledgments

I would like to thank my advisor, Dr. Xuguang Wang, for her guidance and advice throughout the entire course of my research for this thesis. I would also like to thank my committee members, Dr. James Ruppert, for guidance on analyzing and discussing various tropical features, and Dr. Steven Cavallo, for additional support and feedback to further improve my thesis. Further thanks go to Dr. Yongming Wang for assistance in setting up and running GSI with and without multi-resolution ensemble capabilities as well as the FV3GFS. This study was conducted with support from NOAA and an American Meteorological Society Graduate Fellowship. All computational analysis and forecast generation, post processing, and verification utilized resources provided by NOAA's Research and Development High-Performance Computing System Jet machine.

TABLE OF CONTENTS

Abstract

One challenge associated with ensemble-based data assimilation (DA) is the quantitative estimation of the background error covariances. In the current global operational DA system, a reduced resolution background ensemble is adopted, which restricts the ensemble background error covariance from resolving a wide range of scales. A recent study developed a multi-resolution ensemble four-dimensional ensemble-variational (4DEnVar) method and determined that this approach can provide a nearly comparable analysis and forecast to a single high-resolution ensemble approach with the same ensemble size, while substantially lowering the computational cost. This study further develops the multi-resolution ensemble 4DEnVar approach to allow for a flexible number of low- and high-resolution ensemble members as well as different localization length scales between the high- and low-resolution ensembles. The goal of such development is to determine a more effective way of sampling large and small scales to formulate a method aimed at achieving an effective multiscale DA system.

Utilizing the Finite-Volume Cubed Sphere Global Forecast System (FV3GFS), three 4DEnVar experiments with the same computational costs are compared. The first has an 80-member high-resolution background ensemble with single-scale localization. The second and third utilize the multi-resolution ensemble developments. One has a 170-member multi-resolution background ensemble, including 130 low-resolution and 40 high-resolution members. The other has a 204-member multi-resolution background ensemble with 180 low-resolution and 24 high-resolution members. Both multi-resolution ensemble experiments also utilize scale-aware localization radii. Despite having the same costs, the multi-resolution ensemble method decreases tropical cyclone track errors and improves global forecasts for up to five days in lead time. Diagnostics suggest that the multi-resolution ensemble method demonstrates improvement in

analyzing large-scale features. At early lead times, improvements are most apparent in the upper levels in the tropics, while at later lead times, the Southern Hemisphere extratropics demonstrates greater improvement. The reduction of forecast errors in the extratropics at later lead times is partially due to improved analyses of large-scale features in the tropics that shift poleward. With regard to tropical cyclone track errors, the multi-resolution ensemble method can more properly correct the large-scale steering flow, leading to more accurate subsequent track forecasts. Additionally, the multi-resolution ensemble with a greater number of high-resolution members performs better than the multi-resolution ensemble with a greater number of low-resolution members in most metrics, demonstrating the value of both increasing sampling at large scales and retaining substantial information at small scales.

1. Introduction

There has been steady improvement in global numerical weather prediction (NWP) from the latter half of the 20th century through the present (e.g., Hamill 2010; Magnussen 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. Specifically, greater efficiency in computing resources have allowed for higher resolution and greater complexity in forecasts (e.g., Bauer et al. 2015). Improved model dynamics and physics have led to more accurate numerical representations of atmospheric processes (e.g., Magnussen and Källén 2013; Benjamin et al. 2019). Effective utilization of existing and new observations through the advancements in data assimilation (DA) have led to more accurate initial conditions (ICs) for numerical models and therefore more accurate subsequent numerical forecasts (e.g., Magnussen and Källén 2013; Magnusson et al. 2019; Chen et al. 2019).

Recent studies demonstrate the advancements of model dynamics and physics in the US Global Forecast System (GFS). In June of 2019, the dynamical core of the operational GFS was updated from the Global Spectral Model (GSM) to the Finite-Volume Cubed Sphere (FV3; Lin and Rood 1997; Lin 2004; Putman and Lin 2007). FV3 uses a semi-Lagrangian flux form of the momentum equations defined over near-cubical volume grids, while the GSM used a wave form of the momentum equations defined using spherical harmonic basis functions (Lin and Rood 1996; Lin 1997; Putman and Lin 2007; Harris et al. 2021). Testing of the pre-operational FV3-based GFS (or FV3GFS) has shown that the updated dynamical core can match or improve global forecasts (Ji 2016; Magnusson et al. 2019) and improve tropical cyclone forecasts relative to the GSM (Chen et al. 2019). The new GFS also utilizes a new physics package developed by the Geophysical Fluid Dynamics Laboratory (GFDL). The package includes an updated single-

moment 6-category cloud microphysics (Chen and Lin 2011, 2013), which aims to improve upon the forecast of moist processes that could not be effectively parameterized in the previous Zhao-Carr microphysics scheme (Zhao and Carr 1997). Zhou et al. (2019) demonstrated that this updated microphysics scheme improved the global forecast due to its ability to parameterize moist processes more accurately at increased resolution.

Additionally, recent studies have shown the need to improve the GFS ICs (Magnusson et al. 2019; Chen et al. 2019). Magnusson et al. (2019) compared forecasts for the FV3GFS, legacy GFS, and ECMWF using ECMWF and GFS ICs. They found that 500-hPa height forecasts produced by the FV3GFS with ECMWF ICs had lower root square mean errors (RMSEs) than that produced by using GFS ICs on average. They also demonstrated that the difference in ICs played a larger role than model formulation in determining these forecast RMSEs up to eight days in lead time. Chen et al. (2019) used a similar methodology to Magnusson et al. (2019) but with a focus on tropical cyclone forecasting. They determined that the FV3GFS can have more skillful tropical cyclones forecasts relative to the ECMWF model when the same ICs are used. These studies demonstrate the importance of developing and utilizing advance DA techniques to improve the ICs for global numerical weather prediction.

Operational numerical prediction centers utilize various advanced DA techniques. One popular category includes the ensemble-based DA approaches, such as various forms of ensemble Kalman filter (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; Wang et al. 2007b; Wang 2010; Wang et al. 2013; Wang and Lei 2014) methods. These approaches use a finite number of ensemble members to sample the background error covariance (BEC) of the short-term

forecast. These BECs derived from the ensemble are flow-dependent, which is particularly advantageous when compared to other common DA methods, such as pure variational methods, which use static BECs based on climatology (e.g., Fischer et al. 1998; Buehner et al. 2013; Wang et al. 2013; Wang and Lei 2014; Kleist and Ide 2015; Buehner et al. 2015).

Among the ensemble-based DA methods, both the legacy GFS and the FV3GFS use the hybrid 4DEnVar method implemented in the gridpoint statistical interpolation (GSI) system (Wang and Lei 2014; Kleist and Ide 2015). 4DEnVar is a 4D-extension of hybrid 3DEnVar (Wang et al. 2013) and has the benefit of not having to code a tangent linear model compared to 4DVar. 4DEnVar has been shown to improve upon both global- (Buehner et al. 2013; Wang and Lei 2014) and convective-scale (Lu et al. 2017; Gao et al. 2021) forecasts when compared to 3DEnVar due to the 4DEnVar's ability to account for the temporal change of the ensemble BECs within the DA window. Past studies also demonstrate that hybridizing a well-constructed static covariance in EnVar can improve upon a pure EnVar or pure EnKF for both global and convective-scale applications given the ability of the static covariance to alleviate the sampling error and to account for model errors (e.g., Lorenc 2003; Wang et al. 2007a; Wang et al. 2009; Wang and Wang 2021).

The current global 4DEnVar operational system at the National Centers for Environmental Prediction (NCEP) uses a single control member at high-resolution and 80 ensemble members all at the same reduced resolution (JCSDA 2018). Given the limited computational resources, the optimal balance between model resolution and ensemble size in the context of ensemble-based DA, therefore, needs to be examined. On one hand, the use of a high-resolution ensemble may be important in resolving smaller-scale background errors (e.g., Hamill and Whitaker 2005). On the other hand, increasing the ensemble size will reduce the sampling error of the ensemble-estimated BECs. Specifically, the typical ensemble size of ~$O(100)$ is much smaller than the degrees of

freedom of the model (Houtekamer and Zhang 2016). This discrepancy is often characterized by spurious distant correlations (e.g., Hamill 2006). Various studies have shown that increasing the ensemble size lowers the sampling error and creates a more accurate representation of the 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, utilizing 4DEnVar in the legacy GFS. They found that though both provide some benefit, 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 in global numerical prediction. There is an even greater body of literatures examining the tradeoff between ensemble size and model resolution in the ensemble forecast context (Mullen and Buizza 2002; Reynolds et al. 2011; Ma et al. 2012; Raynaud and Bouttier 2017). These studies found that at a given computational cost the impact of increasing the ensemble resolution versus increasing the ensemble size of a low-resolution ensemble depended on the features and forecast lead time being examined. Despite their focus on ensemble forecasts, these results further motivate the need to study the model resolution and ensemble size tradeoff in the ensemble DA context.

In addition to universally increasing resolution of all background ensemble members or directly increasing the size of the entire background ensemble members, studies have also explored the optimal ensemble size and resolution for the ensemble-based DA using a multi-resolution background ensemble approach (MR-ENS, hereafter). In this approach, 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). Their study compared experiments using a MR-ENS LETKF and a single-resolution ensemble LETKF at a similar computational cost for the forecast. They found that the MR-ENS LETKF outperforms the single-resolution ensemble LETKF. However, there were limitations to this study. For example, the study provided comparisons with a constraint only on the forecast computational time, neglecting the computational cost for the DA steps. Additionally, since the tests were conducted in a Lorenz model framework, little insights were given into the mechanisms driving the improvements in the MR-ENS LETKF over its single-resolution counterpart.

Kay and Wang (2020) introduced a MR-ENS approach in GSI 4DEnVar with applications to NCEP GFS. Using the same number of ensemble members, 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 single-resolution ensemble 4DEnVar with 80 low-resolution members with a 181% cost increase. Additionally, they found that forecast errors of the MR-ENS approach were much more comparable to a single-resolution ensemble 4DEnVar with 80 high-resolution members while MR-ENS reduced the computational cost by 40%. Overall, the study demonstrated the potential for utilizing the MR-ENS in 4DEnVar to improve global forecasts. However, there are several limitations in the study of Kay and Wang (2020). First, their comparison did not constrain the computation cost. Second, their study utilized the same number of low- and high-resolution members in the MR-ENS background ensemble, which limited the ability to explore the effect of greater sampling at different scales of interest on the analysis and subsequent forecast. Third, they used the same covariance localization distance, or radius at which an observation can impact analysis error correlations, for the low- and high-resolution ensemble in the MR-ENS configuration, which limits the impact of estimating error covariances at multiple scales.

This study aims to further develop and explore the MR-ENS approach in 4DEnVar. First, this study will further develop 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 (e.g., Rainwater and Hunt 2013). This capability would 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 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. This experiment setup allows to address 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 large, medium, and small scales? Addressing these questions using the further developed MR-ENS 4DEnVar is one step toward achieving the ultimate effective multiscale data assimilation for the next-generation global NWP (e.g. Wang et al. 2021).

The rest of this thesis is organized as follows. Section 2 describes the formulation of the single-resolution hybrid 4DEnVar and the newly extended multi-resolution ensemble hybrid 4DEnVar. Section 3 describes configurations of the single- and multi-resolution ensemble experiments, including DA and model parameters. Results comparing the multi-resolution ensemble 4DEnVar and single high-resolution ensemble 4DEnVar are discussed in section 4. The main conclusions are summarized and potential areas for future work are given in section 5.

2. Formulation of single- and multi-resolution ensemble hybrid 4DEnVar

The following describes the 4DEnVar formulations using the notations of Wang (2010), Wang and Lei (2014), and Kay and Wang (2020).

a. Single-resolution ensemble hybrid 4DEnVar

The hybrid 4DEnVar method described in Wang and Lei (2014) and used in operational systems creates an analysis increment $\mathbf{x}'_t$, or the difference between the analysis and the background forecast, for the control member at times $t = 1, \ldots, L$. The analysis increments are defined as

(1)

$$\mathbf{x}'_t = \mathbf{u}\{(\mathbf{x}'_1)_t + \sum_{k=1}^{K} [\,\boldsymbol{\alpha}_k + (\boldsymbol{x}^e_k)_t]\}$$

where $\mathbf{u}$ interpolates low-resolution grids to high-resolution. If all components are at the same resolution, $\mathbf{u}$ is an identity matrix. $(\mathbf{x}'_1)_t$ is the analysis increment associated with the static error covariance. $\boldsymbol{\alpha}_k$ is the kth augmented control vector (Wang 2010), where k is a member of an ensemble with size K. $(\boldsymbol{x}^e_k)_t$ is the kth background ensemble perturbation normalized by $\sqrt{K-1}$ at time t.

The analysis increments are found by the minimization of a cost function J with respect to augmented control vectors representing the static and ensemble BECs as follows:

(2)

$$\begin{aligned} J[(\mathbf{x}'_1)_t, \boldsymbol{\alpha}] &= \beta_1 J_1 + \beta_e J_e + J_o \\ &= \frac{1}{2}\beta_1 (\mathbf{x}'_1)_t^{\mathrm{T}} \mathbf{B}_1^{-1} (\mathbf{x}'_1)_t + \frac{1}{2}\beta_e (\boldsymbol{\alpha})^{\mathrm{T}} \mathbf{A}^{-1} (\boldsymbol{\alpha}) + \frac{1}{2}\sum_{t=1}^{L} (\mathbf{H}\mathbf{x}'_t - \mathbf{y}^o_t)^{\mathrm{T}}\, \mathbf{R}^{-1} (\mathbf{H}\mathbf{x}'_t - \mathbf{y}^o_t) \end{aligned}$$

The static BEC $\mathbf{B}_1$, as traditionally used in GSI 3DVar, is associated with the analysis increment $(\mathbf{x}'_1)_t$. $\boldsymbol{\alpha}$ is a concatenation of augmented control vectors over all ensemble members and is

constrained by the block-diagonal matrix **A**, which is used to localize the covariance of the ensemble. **A** consists of two parts: $\mathbf{A_h}$ and $\mathbf{A_v}$. $\mathbf{A_h}$ represents the horizontal covariance localization, which is a Gaspari and Cohn (1999) function with a set e-folding distance that varies with the vertical level of the model as in the operational configuration (e.g., Wang and Lei 2014; Kleist and Ide 2015; Lei and Whitaker 2017; Huang and Wang 2018; Kay and Wang 2020). The vertical covariance localization $\mathbf{A_v}$ is applied through a recursive filter transform (Hayden and Purser 1995) and is given as a fixed level-invariant length scale. β_1 and β_e are the weights given to the static and ensemble BECs respectively, such that $1/\beta_1 + 1/\beta_e = 1$ (Wang et al. 2013). **H** is the linearized observation operator, which maps the analysis increment from model space to observation space. $\mathbf{y}_t^o$ is the innovation vector, which is the difference between the observation and background forecast, at time t. **R** is the observation error covariance matrix.

b. Multi-resolution ensemble (MR-ENS) hybrid 4DEnVar

The multi-resolution ensemble (MR-ENS) 4DEnVar method as described in Kay and Wang (2020) extends the hybrid 4DEnVar formulation to incorporate BECs from a multi-resolution ensemble, rather than the single-resolution formulation used in Wang and Lei (2014). The variables representing the low- and high-resolution ensembles will be represented by *L* and *H* respectively. For this formulation, the analysis increments are defined by

(3)

$$\mathbf{x}'_{\mathrm{t}} = (\mathbf{x}'_1)_t + \mathbf{u}\sum_{k=1}^{K_L}[\,\boldsymbol{\alpha}_k^L + (\boldsymbol{x}_k^{eL})_t] + \sum_{k=1}^{K_H}[\,\boldsymbol{\alpha}_k^H + (\boldsymbol{x}_k^{eH})_t]$$

$\boldsymbol{\alpha}_k^L$ and $\boldsymbol{\alpha}_k^H$ are the kth augmented control vectors for the low- and high-resolution ensembles respectively. $(\boldsymbol{x}_k^{eL})_t$ and $(\boldsymbol{x}_k^{eH})_t$ are the kth background ensemble perturbations normalized by $\sqrt{K_L - 1}$ and $\sqrt{K_H - 1}$, where K_L and K_H are the sizes of the low- and high-resolution ensembles.

K_L and K_H may have the same or different values. However, in Kay and Wang (2020), K_L and K_H were both 40.

As with the single-resolution ensemble, the analysis increments are found by minimizing the hybrid cost function as follows:

(4)

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

$\boldsymbol{\alpha^L}$ and $\boldsymbol{\alpha}^H$ are the concatenation of augmented control vectors for the low- and high-resolution ensembles with K_L and K_H members respectively. $\mathbf{A}_L$ and $\mathbf{A}_H$ are used to determine the covariance localization in the horizontal and vertical for the low- and high-resolution ensembles respectively. As with the number of ensemble members, the covariance localization may also vary between the low- and high-resolution ensembles, but in Kay and Wang (2020), the localization was kept the same. β_L and β_H are the weights given to the low- and high-resolution ensemble BECs and are constrained by $1/\beta_1 + 1/\beta_L + 1/\beta_H = 1$.

3. Experimental design

The GSI-based 4DEnVar DA system and FV3GFS forecasts were cycled every six hours for a five-week period—from 25 August 2017 at 1800 UTC to 29 September 2017 at 1200 UTC. This period was chosen, in part, so that a portion of the diagnostics could focus on the substantial number of tropical cyclones that occurred during this month, including multiple high-intensity storms in the Atlantic basin. In all diagnostics, the first week of this period was removed to account for spin-up time, as in Kay and Wang (2020), so that the results were not biased by the choice of the initial ensemble. The system assimilates the NCEP operational global conventional and satellite observations[1] over a six-hour DA window.

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

One 4DEnVar experiment was run at a single resolution, where the background control member and the BECs of the static and single-resolution ensemble were all at the same high-resolution (T766 or ~25 km). These static and ensemble BECs were used to update the control member as in section 2a. As in the current operational 4DEnVar system, 80 ensemble members were used. The static and ensemble BECs were given weights of 12.5% and 87.5% respectively (Lei and Whitaker 2017; Huang and Wang 2018; Kay and Wang 2020; Huang et al. 2021). Level-dependent horizontal covariance localization was utilized by applying a Gaspari and Cohn (1999) localization function with the same localization distances as applied in Kleist and Ide (2015) and Lei and Whitaker (2017). The horizontal *e*-folding distance was set to 350 km from the surface to ~300 hPa (*hl1*), 1000 km between 56 and 14 hPa (*hl2*), and 1300 km from 5 hPa to the model top (*hl3*) with a linear transition between these levels, as shown in the blue line in Fig. 1. In the vertical direction, a constant localization e-folding distance of 0.5 scale heights was applied.

[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

Before updating the ensemble, the high-resolution ensemble was interpolated to low resolution (T382 or ~50 km) due to computational constraints of the ensemble update and for consistency with the MR-ENS experiments, as in Kay and Wang (2020). The ensemble was updated using 4D LETKF (Bishop et al. 2001; Wang and Bishop 2003; Wang et al. 2004; Hunt et al. 2007) in GSI. Covariance localization, given as a level-dependent cutoff distance, was equal to 677.835 km for *hl1*, 2577.320 km for *hl2*, and 3350.515 km for *hl3*, with a constant vertical localization cutoff distance of ~0.773 scale heights. Multiplicative inflation (Whitaker and Hamill 2012) was used as a measure to increase ensemble spread by relaxing the spread of the ensemble after DA (i.e., the posterior ensemble) to 85% of the spread of the background ensemble (i.e., the prior ensemble).

The background forecasts for the single control member as well as the 80-member ensemble were completed at high-resolution (C384, or ~25 km) on a cubed-sphere grid with 64 vertical model levels by the June 2020 operational version of the FV3GFSv15. The configuration of the model-specific parameters is detailed in Zhou et al. (2019). To improve balance conditions 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. A flowchart of the general steps in the experiment is given in Fig. 2a.

b. Multi-resolution (MR) Ensemble Experiments

The multi-resolution (MR) ensemble 4DEnVar experiments were run utilizing the extended MR-ENS methodology as shown in section 2b, where the high-resolution control member was updated using a high-resolution static BEC and contributions to ensemble BECs from a low-resolution (T382 or ~50 km) and high-resolution (T766 or ~25 km) ensemble. Further development from Kay and Wang (2020) allowed for varying ensemble sizes. Two such

configurations, which had near equal computational costs compared with SR-High (Table 1), were examined to test the sensitivity of the ensemble size of the low- and high-resolution ensembles. The first MR 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 MR 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. Similar computational costs between all experiments allowed for a direct examination of analysis and forecast errors. The MR 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. The vertical localization was applied, as in the SR-High experiment, with a constant localization e-folding distance of 0.5 scale heights.

Additional developments to Kay and Wang (2020) made it possible to vary the horizontal localization distance between the low- and high-resolution ensembles. Previous studies have indicated that for a larger ensemble size, the optimal localization radius is larger than for a smaller ensemble size, leading to greater dynamical balance and for additional distant non-spurious correlations to be made (e.g., Houtekamer and Mitchell 2001; Hamill et al. 2001; Ying et al. 2018). For these MR experiments, for the high-resolution ensemble, the level-dependent localization distances (*hl1*, *hl2*, and *hl3*) were set as equal to those described above in the SR-High experiment (blue line in Fig. 1). 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. 1). This change allows for correlations to be present at wider spatial scales.

LETKF updates to the ensemble were all made at low-resolution, as in the SR-High experiment. All LETKF covariance localization and inflation parameters were the same as described in section 3a.

The FV3GFS background forecasts were completed at high-resolution (C384 or ~25 km) for the control member and high-resolution ensemble and at low-resolution (C192 or ~50 km) for the low-resolution ensemble. The flowchart for the MR experiments is shown in Fig. 2b.

4. Results

a. Multiscale analysis

One way to determine the multiscale impact of the extended MR-ENS and single-resolution ensemble 4DEnVar techniques on the analysis is to decompose analysis increments into spectral space (e.g., Kay and Wang 2020; Huang et al. 2021). This decomposition will show how much of an adjustment the DA method made to the background forecast at each scale when calculating the analysis. For each experiment at each analysis time, the analysis increments were calculated by subtracting the background forecast from the analysis at a given level. The analysis increments were then decomposed into spectral space using spherical harmonics and averaged over all cycles. The 500-hPa analysis increment power spectrums for temperature and zonal wind are shown in Fig. 3. For both temperature and zonal wind, the MR170 experiment has larger increments for larger atmospheric scales (wavenumbers less than ~150 or wavelengths greater than ~200 km) when compared with the SR-High experiment. For smaller scales (wavenumbers greater than ~150 or wavelengths less than ~200 km), the MR170 experiment has similar or slightly smaller analysis increments.

These results are consistent with the MR and SR-High methodologies. The MR170 experiment has a broader localization scale with an e-folding distance of 700 km applied to its low-resolution ensemble BECs, while SR-High has an e-folding localization distance of 350 km for its single-resolution ensemble BECs. This difference implies that using the extended MR-ENS technique allows for larger adjustments to be made to the background ensemble at a further distance from the observation point, thereby indicating larger analysis increments at large scales. Huang et al. (2021, their Fig. 6) shows a very similar result where two of their experiments at the same ensemble size and resolution have vertically invariant horizontal localization radii with e-

folding distances of 300 and 1000 km. Comparing these experiments, the 1000-km experiment had larger analysis increments at all scales.

For smaller scales, even though SR-High and MR170 have the same high-resolution localization distance, SR-High has a greater number of high-resolution ensemble members (80 members) when compared to MR170 (40 members). SR-High also has 87.5% of the weight on the ensemble portion of the cost function, whereas MR170 has a weight of 43.75% applied to the high-resolution ensemble portion of the cost function. These differences in methodology create a lower effective resolution as described in Kay and Wang (2020) and lead to smaller adjustments at smaller scales in accordance with the lesser weight on the high-resolution ensemble BECs in the MR170 experiment. The smaller adjustments are reflected in the lower power in analysis increments at larger wavenumbers.

Another way of examining the multiscale impact of the different DA methods on the analysis is to complete a wavelet diagnostic on the analysis increments for the experiments (Torrence and Compo 1998). A wavelet analysis, in this case, can partially diagnose both the location and spatial scale of decomposed analysis increments. The kinetic energy analysis 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 all 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 being affected.

For the largest scales, there is a general increase in kinetic energy analysis increments globally for the MR170 experiment (Fig. 4a). This increase is mainly driven by features in the SH region (Fig. 4c), indicating that at the largest scales, the extended MR-ENS technique has a greater impact on the analysis in this region. At the smallest scales, globally, the SR-High experiment has substantially greater analysis increment power at some longitudes—namely between 15°W and 90°W. At other longitudes, there is little difference between the two experiments. The largest and smallest scales display consistency with the results discussed above in the spectral decomposition of analysis increments at 500 hPa.

At medium scales, the experiments produced mixed results with regards to their effect on the kinetic energy analysis increments. For most longitudes, the MR170 experiment still has greater analysis increment power. However, between about 120°W and 0°, the analysis increment power for the SR-High experiment is substantially increased. As shown in Fig. 4b,d, this increase is primarily driven by the NH and TR in the areas associated with the Atlantic Ocean, the Caribbean Sea, the Gulf of Mexico, and the United States East Coast. Contrarily, in the SH, MR170 has the greater analysis increment power between the two experiments in the Atlantic region.

As with the large and small scales, the SH results at all longitudes as well as the NH and TR results outside of the Atlantic region for medium scales are not unexpected as they are consistent with Fig. 3 in that the larger localization distance is leading to larger analysis increments at large and medium scales. However, the larger kinetic energy analysis increment power for SR-High in medium scales in the Atlantic is antithetical to this notion. To determine what physical

processes are associated with this result, instead of averaging the wavelets over latitude, as in Fig. 4, the wavelets were instead averaged over all levels, all cycles, and small- (wavelet scales 1-15 or wavelengths less than approximately 350 km) and medium-scales (wavelet scales 16-30 or approximately wavelengths between 350 and 5000 km). Then, the difference between MR170 and SR-High was taken and plotted in Fig. 5.

Medium scales display a minimum in the difference in analysis increment power in the subtropical and northern tropical Atlantic (Fig. 5a). Tropical easterly wave tracks that were observed at 850 hPa during September 2017 as calculated by Hollis (2021) are also overlaid in Fig. 5a. Tropical easterly waves, of which the most well-known subset is African easterly waves, are westward propagating synoptic-scale waves in the tropics, often associated with areas of convection (e.g., Carlson 1969). These waves are frequently precursors to tropical cyclones (TCs), especially in the Atlantic basin (e.g., Landsea and Gray 1992). Though the waves themselves have typical wavelengths of 2000 to 4000 km (e.g., Carlson 1969), the most active region of moist convection spans a shorter distance. Hollis (2021) found that a 500-km radius was sufficient to capture the area of intense precipitation associated with tropical easterly waves, which is an area consistent with the medium scale defined above. Comparing the wavelet power to the wave locations, the greater analysis increment power of the SR-High analysis at medium scales is substantially collocated with the high density of tropical easterly waves in the Atlantic.

To further confirm that these types of high-intensity cyclonic features are dominating the increase in kinetic energy analysis increments for SR-High, the small-scale analysis increments were also plotted in Fig. 5b, which generally shows an area of high-magnitude negative values in the northern tropical Atlantic and near the United States' East Coast. Comparing these differences to the Atlantic hurricanes during that month, three hurricane tracks fall within that area. These are

Hurricanes Irma, Jose, and Maria, which were the three strongest hurricanes during the 2017 season, peaking at a Category 4 for Jose and Category 5 for Irma and Maria. It would be reasonable that these storms—with very strong winds over small scales in their powerful inner-core and rainband regions—would have larger analysis increments at small scales for a DA method more apt to resolve small-scale features.

In features for which the most intense impacts are dominated by the mesoscale, such as in TCs and tropical easterly waves, SR-High has the advantage of making larger adjustments in these scales. These adjustments are due to the greater number of high-resolution ensemble members and larger weight on the tighter localization scale, which has been shown to be beneficial for smaller-scale features (e.g., Zhang et al. 2009). In more synoptically forced regions, however, as is the case in the SH in the austral late winter/early spring during which this study occurs, MR170 still makes larger corrections in the medium to large scales, as would be expected with a larger number of low-resolution members and a broader localization scale applied to these members.

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.

b. Short-term forecast errors and ensemble spread

There are several methods that, when employed, can determine if the greater analysis increments, as examined in the previous section, translate into a better forecast. One method is to compare the forecast to the observations. For this study, the difference between the 6-h forecast in observation space and the rawinsonde observations of temperature and wind speed were calculated for each cycle. The root square mean fit (RMSF) of the forecast to the observations for

all cycles was then calculated and averaged over vertical levels in increments of 100 hPa. The RMSF was broken down into the same four regions as described in section 4a: 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 with a 95% confidence interval, were plotted. For both the temperature and wind speed, MR170 has a better fit to the observations when compared with SR-High for all vertical levels (Fig. 6). The improvement is most apparent in the upper levels of the TR region for both the temperature and wind speed.

Another aspect that it is important to examine when using an ensemble-based DA method is the relationship between the error and the spread of the ensemble. Ideally, the spread of the ensemble, which is a measure of the uncertainty of the forecast, would be representative of the ensemble forecast errors. However, ensemble-based DA systems often have ensemble spread that is relatively too small. This deficiency can lead to an issue called 'filter divergence,' where over several cycles, the observations become underweighted due to overconfidence in the ensemble forecast (e.g., Houtekamer and Mitchell 1998; Whitaker and Hamill 2002). In general, the total RMSF of the 6-h forecast to temperature and wind observations in Fig. 7 shows that the temperature errors are largest at the surface and upper levels for all regions, while wind speed 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, the ensemble is shown to be underdispersive for both the temperature and the wind speed. However, this difference between the RMSF and spread is smaller for the MR170 experiment than for the SR-High experiment.

Fig. 8 further demonstrates how the ensemble spread is increased in the MR170 experiment for most vertical levels. This increase is most apparent in the SH, especially in levels that are

typically associated with substantial temporal and spatial variance in temperature (i.e., near the surface) and wind speed (i.e., near jet-level). The reason behind this increase may not be apparent, at first, since DA methods using higher resolution, 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, other factors are likely greater contributors to the increased spread for the MR170 experiment. The increase in spread, in part, might be due to the increase in the total number of ensemble members in the MR170 experiment. Increased ensemble size typically leads to a decrease in sampling error, which could lead to a spread more representative 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 (e.g., frontal regions at the surface or jet boundaries). This notion can be demonstrated by examining the 200-hPa wind speed and increased wind speed ensemble spread in the SH. Over the experiment period, the regions associated with the greatest increase in spread consistently corresponds with the strongest jet regions (not shown).

c. Forecast errors at longer lead times

Further analysis of forecast errors was conducted by running a 120-h FV3GFS deterministic forecast for each experiment initialized at each analysis time throughout the experiment period. The root mean square error (RMSE) of the forecasts was calculated by taking the root mean square of the difference of the experiment forecast at each forecast lead time and the ECMWF Reanalysis-Interim (ERA-Interim; Berrisford et al. 2011) one-degree data for the same time. The temporally averaged difference in the RMSE for temperature and wind speed between the MR170 and the SR-High experiments for 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 (Fig. 9a,b). The magnitude of this improvement generally grows for later lead times with the greatest improvement predominately near jet level. The differences in the RMSE for the NH and SH regions mainly resemble the global pattern (Fig. 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. The TR region also shows general improvement for the MR170 experiment with the exception of the temperature at the lowest levels for most lead times (Fig. 9g,h). In contrast to the other regions, the largest improvement is at the analysis and early lead times with the forecast errors becoming more similar between the two experiments at later lead times.

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

(5)

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

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

The difference between the spectrally decomposed errors for the two experiments averaged over the duration of the experiment is shown in Fig. 10. At analysis time, the MR170 experiment, on average, shows improvement in energy errors at nearly all wavenumbers. These improvements reach a first maximum at around 24 hours between wavenumbers 5 and 25 (wavelengths of

approximately 8000-1600 km). This peak disperses into larger and smaller wavenumbers, and a second maximum occurs at similar wavenumbers starting right before 72 hours. From 72 to 120 hours, 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 errors between the two experiments, one case, which was representative of the errors shown in Figs. 9,10, was chosen. The case had an analysis time of 0000 UTC on 12 September 2017, and the 250-hPa level was chosen to be analyzed since pressure levels between 200 hPa and 300 hPa typically demonstrated the greatest improvement in Fig. 9. To determine specifically what was occurring at the peaks in energy 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 this result to filter for 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. Fig. 12 shows the locations of the 5% greatest improvement for the MR170 experiment, filtered for the same wavenumbers in 24-h increments. The figure also displays the associated 250-hPa wind speed map at the same times.

At analysis time, there is peak in the tropics, especially near the equator, of the difference in energy errors, indicating improvement for MR170 in this region (Fig. 11). The largest area of improvement is collocated with a region of high wind speeds near parallel to the western coast of Africa that connects with the subtropical jet at around 20°S (Fig. 12a). The northern portion of this high-wind speed feature is consistent with the location of the western portion of the Tropical Easterly Jet (TEJ; Koteswaram 1958). The MR170 analysis displays substantial improvement both within the TEJ and in the area where it connects with the subtropical jet. By 24 hours, some

of the initial improvement in the tropics continues to interact with the SH subtropical jet, expanding and intensifying the areas of improvement in the tropics and SH subtropics (Figs. 11,12b). Between 24 and 72 hours, locally generated errors in the NH and SH extratropics begin to become evident (Fig. 12b-d). These errors are largely collocated with regions of the polar jet in both hemispheres. Additionally, during this time period, areas of interaction between the polar and subtropical jets in the SH tend to exhibit more substantial growth of the differences in errors. After 72 hours, the error differences continue to grow in both hemispheres but more so in the SH than the NH. This growth is especially apparent in the highly baroclinic region of the polar jet in the SH, especially in the areas to the southwest of Australia, south of Africa, and to the southeast of South America (Fig. 12e,f).

The 12 September case study confirms several details of Figs. 9,10. For Fig. 9, it confirms that the largest differences in upper-level temperatures and wind speeds between the two experiments are in the tropics at the analysis time and early lead times. The case study also confirms that as the difference in the errors become less centered in the tropics at later lead times, error differences are simultaneously shifting to the extratropics, most notably in the SH. Since the case study focused on wavenumbers 5 through 25, it emphasizes the impact and importance of these wavenumbers on the forecast as shown in Fig. 10.

The case study also elucidates a few aspects not immediately evident in Figs. 9,10. It demonstrates what sort of features in the tropics are originally analyzed better in the MR170 experiment. In this case, the most notable feature is the TEJ and its connection with the subtropical jet. The case also highlights the importance of the interaction between the tropics, subtropics, and extratropics in terms of error growth. In the NH, the jet streams are mainly weaker and isolated from the tropical region as would be expected in the late summer/early spring. Therefore, error

growth in this region depends mainly on errors that were 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, which are already larger than in the NH due to the strength of the jet in the SH late winter/early spring, are further compounded by the errors at analysis and early lead times in the tropics and subtropics. It should be noted, however, that this study only reviews a single case in depth. A deeper examination of several cases over multiple seasons, would yield more fruitful insights into the origination, propagation, and seasonality of forecast error growth.

Figs. 9,10 demonstrate that MR170 improves the analysis for nearly all scales and pressure levels on average during the experiment period. This initial improvement leads to a better subsequent forecast, with the largest improvements in the large-scale upper-levels tropics at early lead times and the large-scale upper-level extratropics at later lead times (Figs. 9,11,12). The improvements in the analyses and forecasts, as they are most apparent at large scales, can be attributed to two interconnected factors: (1) the larger ensemble size in the MR170 experiment due to the large number of low-resolution members; (2) the larger localization radius for the low-resolution ensemble. The larger ensemble size leads to a more accurate sampling of the background error correlations at all scales with wavelengths larger than ~100 km, (i.e., twice the approximate grid spacing of the low-resolution ensemble). The increased sampling for the low-resolution ensemble also motivates the necessity of using a larger localization distance. The use of a larger localization radius for the low-resolution ensemble in the MR170 experiment allows the observations to make corrections at longer distances (i.e., those more relevant to the synoptic scale) to a greater degree than in the SR-High experiment, which uses single-scale localization. Both the

greater sampling and larger localization distance augment the ability to create an analysis with greater emphasis on correcting larger scales. The improvements in both the analysis and forecast provide further evidence, in addition to that provided by previous studies (e.g., Zhang et al. 2009; Miyoshi and Kondo 2013; Huang et al. 2021; Wang et al. 2021), that DA methods focused on further correcting the most relevant scales are necessary to improve upon corresponding DA techniques that are scale unaware.

d. Tropical cyclone track errors

In addition to examining global forecasts, the impact of the extended MR-ENS methodology on tropical cyclone (TC) track 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 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 and the "best track" data as determined by the National Hurricane Center[2] for the Atlantic and East Pacific basins and the Joint Typhoon Warning Center[3] for the West Pacific basin. Track errors were averaged over all 15 storms for each experiment and lead time. Statistical significance for differences between MR and SR-High track errors was calculated by using a paired *t*-test with a 95% confidence interval.

In Fig. 13a, the percent difference in temporally averaged track errors for MR experiments with SR-High used as the reference is shown. For all forecast hours after the analysis time and up to 120 hours, the MR170 experiment shows a statistically significant decrease in TC track errors.

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

The largest improvements occur between 12 and 54 hours where the errors decrease by ~20%; however, MR170 has an improvement in track errors exceeding 10% for all lead times between 6 and 114 hours. The percentage of track errors that are decreased compared with SR-High are shown in Fig. 13b, where a 100% track error decrease would be indicating that the MR experiment has smaller track errors at a certain lead time for all tracks in all 6-h cycles, while 50% would signify an equal likelihood of MR 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 hours.

The combination of results from Fig. 13a,b demonstrates that in the 0-6-h timeframe, there are fewer total tracks that are improved both at and immediately after analysis time for the MR170 experiment; however, the tracks that do have smaller errors are improved by a larger margin on average. This result can also explain why the percentage differences between track errors are larger at 12–54-h lead times when compared to 0-6-h lead times. As discussed in section 4a, the SR-High experiment has larger kinetic energy analysis increment power, especially at mid- 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 to 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, both the necessity of 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—have some importance. At later lead times, it is reasonable that

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 substantial improvement (improvements exceeding 50%) 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 on 24 September 2017. This time occurred soon after the re-intensification of Lee into hurricane strength and was also a period of substantial track uncertainty in 2017 operational models (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). In this study, the large-scale deep-layer environmental flow in the basin surrounding Lee was determined by first calculating a pressure-weighted average of the zonal and meridional winds at each grid point from 850-300 hPa (e.g., Wu and Chen 2016). Then, the effects of the storm itself were removed by filtering the winds to include only large scales (approximately wavenumbers 0-10). These large-scale steering flow winds as well as 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 plotted along with the track positions in Fig. 14.

At the analysis time, in the area south of and immediately surrounding the center of the storm, there is improvement for the MR170 experiment in the large-scale zonal winds associated with the northern portion of a large-scale anticyclonic region (Fig. 14a). As time progresses, this large-scale anticyclone propagates northwestward toward Lee, shifting the large-scale direction of the winds in the vicinity of the storm center from westerly to easterly in the first 36 hours (Fig.

14b-d). The large-scale zonal wind errors associated with the anticyclonic region also remain smaller for the MR170 experiment up to 36 hours. Given that the track of Lee during the first 36 hours is largely zonal (an eastward track in the first 12 hours and a westward track from 18 to 36 hours) and given that this corresponds well with the position of the anticyclone, it is reasonable that the differences in Lee's track for the first 36 hours for the two experiments can be attributed to the improvements in the analysis of the large-scale zonal wind speed. These improvements can be connected to the ability of the extended MR-ENS technique to better correct 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

The MR170 experiment demonstrated the potential for the extended MR-ENS methodology to improve upon the analysis and forecast when compared to the SR-High methodology. In order to further test the sensitivity of the ensemble size and composition, a second MR experiment with 180 low-resolution members and 24 high-resolution members (MR204) was compared with SR-High and MR170.

1) ANALYSIS INCREMENTS

The MR204 500-hPa analysis increments decomposed into spectral space show a very similar pattern to that of the MR170 experiment for both the temperature and the zonal wind (Fig. 3). Both MR experiments show a substantial increase in the analysis increments over SR-High at larger spatial scales as well as similar or slightly smaller analysis increments at smaller scales. Since the two MR experiments both use the same localization distances and only vary the ensemble size and composition, this further confirms that the differences in the analysis increments of the SR-High and MR experiments at large scales is resultant of the differences in the larger localization

radius associated with the low-resolution ensemble in the MR experiments. At smaller scales, the localization associated with the high-resolution ensemble and the weights of the high-resolution ensemble BECs do not vary between MR experiments. Therefore, it is reasonable that the small-scale analysis increments are also very similar between the MR experiments.

2) FORECAST ERRORS

There is a notable improvement in the RMSF of the 6-h global forecast to the rawinsonde observations for MR204 when compared to SR-High for both the temperature and wind speed at all vertical levels (Fig. 15a,b). However, the MR204 experiment slightly degrades the accuracy of the 6-h forecast when compared with the MR170 experiment (Fig. 15c,d). Similarly, at longer lead times, MR204 improves the RMSE compared to SR-High, demonstrating that the improvement in the global analysis leads to a better forecast (Fig. 16). If the same comparison is conducted between MR170 and MR204, it is found that the MR170 experiment improves the global forecast over MR204 at most vertical levels and lead times (Fig. 17). It should be noted, though, that there are a few levels and lead times with significant degradation for the MR170 experiment, such as for temperature at lower levels in the SH prior to 48 hours (Fig. 17e). For TC track forecasts, MR204 has smaller track errors compared with SR-High up to 120 hours with statistically significant differences up to 36 hours (Fig. 13a). Also, between 24 and 108 hours MR204 improves the track forecast more than 50% of the time when compared with SR-High (Fig. 13b). For nearly all lead times, however, there is a degradation in the track forecast for MR204 when compared to MR170, which is consistent with the global forecast result.

In section 4c, it was conjectured that the MR170 forecast was improved over SR-High, at least in part, due to the increase in the ensemble size from the addition of several low-resolution members. Following this logic, it would be reasonable to assume that adding an even larger number

of low-resolution members, as in the MR204 experiment, would result in further forecast improvements over MR170, but this is not the case. Instead, the lack of additional high-resolution members leads to a loss of information at smaller spatial scales and an overall degradation of the forecast. The necessity of high-resolution ensemble members is a result that is more consistent with Lei and Whitaker (2017), which determined that a high-resolution ensemble with smaller size provided a better forecast than a low-resolution ensemble with larger size at the same computational cost.

Overall, both MR experiments improve upon SR-High. However, MR170 shows greater improvement than MR204, which demonstrates that there may be some optimal value for the low- and high-resolution ensemble sizes in a multi-resolution ensemble. In other words, forecast errors can be minimized using an MR-ENS approach by finding a balance between ensemble size and ensemble resolution.

5. Conclusions

a. Summary

This study utilized a multi-resolution ensemble hybrid 4DEnVar that expands upon and further develops the work of Kay and Wang (2020). The extended MR-ENS method allows for a varying number of members of the low- and high-resolution ensemble members as well as different localization length scales for the low- and high-resolution ensembles. One of the goals in further developing the technique is so that this method may be used as one component of an effective multiscale data assimilation system for next-generation global NWP.

In this study, the extended MR-ENS technique was applied in GSI with 6-h cycling with the FV3GFS for a five-week period. Two MR 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. Both experiments had varying horizontal localization length scales with ensemble resolution to allow for the localization distance to be reflective of larger and smaller scales. The high-resolution ensemble had a 350-km horizontal localization e-folding distance, while the low-resolution ensemble had a 700-km horizontal localization e-folding distance. The impact on the analysis and forecast errors of these experiments were compared to a single-resolution hybrid 4DEnVar experiment, SR-High, at the same computational cost. SR-High had 80 high-resolution ensemble members and single-scale localization with a 350-km horizontal e-folding distance.

The extended MR-ENS technique generally corrected a wider range of meteorological scales in the analysis increments. Globally, larger analysis increment power was present for MR experiments at large scales. At the smallest scales, MR experiments exhibited slightly smaller increment power. Based on the differences between the MR and SR-High experiments in

horizontal localization and weight on the high-resolution ensemble, this result is consistent with previous literature (e.g., Huang et al. 2021). In medium scales, the extended MR-ENS technique more substantially corrected the background forecast in most regions, most notably the SH, where large-scale forcing, such as that associated with the extratropical cyclones and jet streams, 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 (four of which were categorized as major hurricanes) and a substantial number of tropical easterly waves.

Larger analysis increment power at most scales for the MR experiments tended to lead to an overall improvement in the analysis and 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 for all regions, which demonstrates the importance of having a DA method that is able to correct these scales. Improvements were especially evident in the upper levels in the tropics at early lead times (prior to 48 hours) and the SH at later lead times. The SH improvements had contributions both from improved analyses at large scales in the SH and from improvements sourced in the tropics. 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. This result demonstrates that it is important to create an analysis that can both correct the large-scale features themselves as well as allow for the more multiscale interaction between these features. Yang et al. (2017) had a similar conclusion in their study, which utilized multiscale covariance localization to examine the impact on forecasting a heavy rainfall event.

As with global forecasts, tropical cyclone track forecasts were generally improved for the MR 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 the 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 average 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. The slight degradation in MR from 0-6 hours compared to other lead times shows correspondence with the larger analysis increments for SR-High in TC basins during the experiment period. The correspondence may indicate that there are still applications for which having greater information at high-resolution via more high-resolution ensemble members with smaller localization length scales is preferable to having greater sampling at low-resolution with longer localization length scales. Notably, this result is most applicable for very short lead times in features with high intensity over small scales. However, at longer lead times, more substantial corrections in the larger scales still seem to provide better forecasts for the propagation of these same systems.

Two MR configurations were compared to determine the sensitivity of the analysis and forecast to varied ensemble compositions. Both MR configurations improved upon the SR-High analysis and subsequent forecast. This improvement further demonstrates the importance of correcting the large scale in a global NWP forecast through reducing sampling error at these scales by using additional low-resolution ensemble members and increasing the horizontal localization radius. Though it could be speculated that adding additional low-resolution members would further improve this result, the MR 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.

b. Future work

This study examined the impact of using multi-resolution capabilities with variable ensemble size to put different emphasis on large or small scales. This method is only one aspect of building an effective multiscale DA system. Additional aspects include using scale-aware covariance localization in both the horizontal and vertical directions. Scale-dependent localization (SDL) is a horizontal covariance localization method that was developed for EnVar applications (Buehner and Shlyaeva 2015). SDL decomposes ensemble perturbations into overlapping wavebands/scales and simultaneously updates these scales using localization radii appropriate for each scale. Huang et al. (2021) further developed and implemented SDL in a GSI-based 4DEnVar framework. Their work found that the use of SDL leads to improvements in global forecasts when compared with scale-unaware covariance localization methods. In the vertical direction, the current operational system at NCEP still uses scale- and level-invariant covariance localization, even as vertical resolution continues to increase (Yang 2020). Significant work is still needed in order to develop effective scale-aware vertical covariance localization. However, once developed, scale-aware vertical localization may be combined with the SDL capabilities, described in Huang et al. (2021), as well as the methodologies of the multi-resolution ensemble DA with variable ensemble size described in this study. Additional developments will be needed to ensure that all three capabilities are compatible with each other.

Also, the current operational version of the GFS (i.e., version 16) is the final major version that will use the GSI framework (Thomas 2022). The transition to the usage of the Joint Effort for

Data Assimilation Integration (JEDI) as the operational DA framework is currently scheduled to be concurrent with the release of the GFSv17 in 2024. Therefore, additional efforts are needed to make the methodologies described in this study as well as other studies addressing multiscale DA compatible with JEDI to aid in effective research-to-operations transitions.

References

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Monthly Weather Review*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

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. *Meteorological Monographs*, **59**, 13.1-13.67, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1.

Berrisford, P., and Coauthors, 2011: The ERA-Interim archive Version 2.0. 23.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, **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 (AL142017).

Bloom, S. C., L. L. Takacs, A. M. da Silva, and D. Ledvina, 1996: Data Assimilation Using Incremental Analysis Updates. *Monthly Weather Review*, **124**, 1256–1271, https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2.

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

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

Buehner, M., and Coauthors, 2015: Implementation of Deterministic Weather Forecasting Systems Based on Ensemble–Variational Data Assimilation at Environment Canada. Part I: The Global System. *Monthly Weather Review*, **143**, 2532–2559, https://doi.org/10.1175/MWR-D-14-00354.1.

Carlson, T. N., 1969: SOME REMARKS ON AFRICAN DISTURBANCES AND THEIR PROGRESS OVER THE TROPICAL ATLANTIC. *Monthly Weather Review*, **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. *Monthly Weather Review*, **110**, 1354–1374, https://doi.org/10.1175/1520-0493(1982)110<1354:TCMASF>2.0.CO;2.

Chen, J.-H., and S.-J. Lin, 2011: The remarkable predictability of inter-annual variability of Atlantic hurricanes during the past decade. *Geophysical Research Letters*, **38**, https://doi.org/10.1029/2011GL047629.

——, and ——, 2013: Seasonal Predictions of Tropical Cyclones Using a 25-km-Resolution General Circulation Model. *Journal of Climate*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

——, and Coauthors, 2019: Advancements in Hurricane Prediction With NOAA's Next-Generation Forecast System. *Geophysical Research Letters*, **46**, 4495–4501, https://doi.org/10.1029/2019GL082410.

Domingues, M. O., O. Mendes, and A. M. da Costa, 2005: On wavelet techniques in atmospheric sciences. *Advances in Space Research*, **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. *Journal of the Atmospheric Sciences*, **56**, 1627–1648, https://doi.org/10.1175/1520-0469(1999)056<1627:SVPGIA>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**, 10143, https://doi.org/10.1029/94JC00572.

Fischer, C., A. Joly, and F. Lalaurette, 1998: Error growth and Kalman filtering within an idealized baroclinic flow. *Tellus A: Dynamic Meteorology and Oceanography*, **50**, 596–615, https://doi.org/10.3402/tellusa.v50i5.14561.

Gao, S., N. Du, J. Min, and H. Yu, 2021: Impact of assimilating radar data using a hybrid 4DEnVar approach on prediction of convective events. *Tellus A: Dynamic Meteorology and Oceanography*, **73**, 1–19, https://doi.org/10.1080/16000870.2021.1903770.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **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. *Journal of Applied Meteorology and Climatology*, **15**, 1252–1264, https://doi.org/10.1175/1520-0450(1976)015<1252:TCMASP>2.0.CO;2.

Hamill, T. M., 2006: Ensemble-based atmospheric data assimilation. *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 124–156.

——, and C. Snyder, 2000: A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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.

Hamrud, M., M. Bonavita, and L. Isaksen, 2015: EnKF and Hybrid Gain Ensemble Data Assimilation. Part I: EnKF Implementation. *Monthly Weather Review*, **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.

Hayden, C. M., and R. J. Purser, 1995: Recursive Filter Objective Analysis of Meteorological Fields: Applications to NESDIS Operational Processing. *Journal of Applied Meteorology and Climatology*, **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.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, **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. *Monthly Weather Review*, **129**, 123–137, https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

——, and F. Zhang, 2016: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. *Monthly Weather Review*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

——, X. Deng, H. L. Mitchell, S.-J. Baek, and N. Gagnon, 2014: Higher Resolution in an Operational Ensemble Kalman Filter. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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: Nonlinear Phenomena*, **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). 34 pp., https://doi.org/10.25923/JW00-R987.

Ji, M., 2016: *Dynamical Core Evaluation Test Report for NOAA's Next Generation Global Prediction System (NGGPS)*. NOAA, https://www.weather.gov/media/sti/nggps/NGGPS%20Dycore%20Phase%202%20Test%20Report%20website.pdf.

Kay, J., and X. Wang, 2020: A Multiresolution Ensemble Hybrid 4DEnVar for Global Numerical Prediction. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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/10.1111/j.2153-3490.1958.tb01984.x.

Landsea, C. W., and W. M. Gray, 1992: The Strong Association between Western Sahelian Monsoon Rainfall and Intense Atlantic Hurricanes. *Journal of Climate*, **5**, 435–453, https://doi.org/10.1175/1520-0442(1992)005<0435:TSABWS>2.0.CO;2.

Lei, L., and J. S. Whitaker, 2017: Evaluating the trade-offs between ensemble size and ensemble resolution in an ensemble-variational data assimilation system. *Journal of Advances in Modeling Earth Systems*, **9**, 781–789, https://doi.org/10.1002/2016MS000864.

Lin, S.-J., 2004: A "Vertically Lagrangian" Finite-Volume Dynamical Core for Global Models. *Monthly Weather Review*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

——, and R. B. Rood, 1996: Multidimensional Flux-Form Semi-Lagrangian Transport Schemes. *Mon. Wea. Rev.*, **124**, 2046–2070, https://doi.org/10.1175/1520-0493(1996)124<2046:MFFSLT>2.0.CO;2.

——, and ——, 1997: An explicit flux-form semi-lagrangian shallow-water model on the sphere. *Q.J Royal Met. Soc.*, **123**, 2477–2498, https://doi.org/10.1002/qj.49712354416.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, **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. *Monthly Weather Review*, **143**, 212–229, https://doi.org/10.1175/MWR-D-14-00195.1.

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

Lu, X., X. Wang, M. Tong, and V. Tallapragada, 2017: GSI-Based, Continuously Cycled, Dual-Resolution Hybrid Ensemble–Variational Data Assimilation System for HWRF: System Description and Experiments with Edouard (2014). *Monthly Weather Review*, **145**, 4877–4898, https://doi.org/10.1175/MWR-D-17-0068.1.

Ma, J., Y. Zhu, R. Wobus, and P. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**, 782–794, https://doi.org/10.1007/s00376-012-1249-y.

Magnusson, L., and E. Källén, 2013: Factors Influencing Skill Improvements in the ECMWF Forecasting System. *Monthly Weather Review*, **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. *Quarterly Journal of the Royal Meteorological Society*, **145**, 2085–2100, https://doi.org/10.1002/qj.3545.

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

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. *Geophysical Research Letters*, **41**, 5264–5271, https://doi.org/10.1002/2014GL060863.

Mullen, S. L., and R. Buizza, 2002: The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System. *Weather and Forecasting*, **17**, 173–191, https://doi.org/10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2.

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

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

Rainwater, S., and B. Hunt, 2013: Mixed-Resolution Ensemble Data Assimilation. *Monthly Weather Review*, **141**, 3007–3021, https://doi.org/10.1175/MWR-D-12-00234.1.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, **143**, 3037–3047, https://doi.org/10.1002/qj.3159.

Reynolds, C. A., J. G. McLay, J. S. Goerss, E. A. Serra, D. Hodyss, and C. R. Sampson, 2011: Impact of Resolution and Design on the U.S. Navy Global Ensemble Performance in the Tropics. *Monthly Weather Review*, **139**, 2145–2155, https://doi.org/10.1175/2011MWR3546.1.

Thomas, C. A., 2022: Future Plans for NCEP Global Data Assimilation: 2023 and Beyond.

Torrence, C., and G. P. Compo, 1998: A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, **79**, 61–78, https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Wang, X., 2010: Incorporating Ensemble Covariance in the Gridpoint Statistical Interpolation Variational Minimization: A Mathematical Framework. *Monthly Weather Review*, **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. *Journal of the Atmospheric Sciences*, **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. *Monthly Weather Review*, **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? *Monthly Weather Review*, **132**, 1590–1605, https://doi.org/10.1175/1520-0493(2004)132<1590:WIBAEO>2.0.CO;2.

——, 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. *Monthly Weather Review*, **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. *Monthly Weather Review*, **135**, 222–227, https://doi.org/10.1175/MWR3282.1.

——, T. M. Hamill, J. S. Whitaker, and C. H. Bishop, 2009: A Comparison of the Hybrid and EnSRF Analysis Schemes in the Presence of Model Errors due to Unresolved Scales. *Monthly Weather Review*, **137**, 3219–3232, https://doi.org/10.1175/2009MWR2923.1.

——, D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-Based Ensemble–Variational Hybrid Data Assimilation for NCEP Global Forecast System: Single-Resolution Experiments. *Monthly Weather Review*, **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). *Monthly Weather Review*, **149**, 605–622, https://doi.org/10.1175/MWR-D-20-0290.1.

Wang, Y., and X. Wang, 2021: Development of Convective-Scale Static Background Error Covariance within GSI-Based Hybrid EnVar System for Direct Radar Reflectivity Data

Assimilation. *Monthly Weather Review*, **149**, 2713–2736, https://doi.org/10.1175/MWR-D-20-0215.1.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Monthly Weather Review*, **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. *Monthly Weather Review*, **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. *Atmospheric Chemistry and Physics*, **16**, 14925–14936, https://doi.org/10.5194/acp-16-14925-2016.

Yang, F., 2020: GFS Development and Transition to Operations.

Yang, S.-C., S.-H. Chen, K. Kondo, T. Miyoshi, Y.-C. Liou, Y.-L. Teng, and H.-L. Chang, 2017: Multilocalization data assimilation for predicting heavy precipitation associated with a multiscale weather system. *Journal of Advances in Modeling Earth Systems*, **9**, 1684–1702, https://doi.org/10.1002/2017MS001009.

Ying, Y., F. Zhang, and J. L. Anderson, 2018: On the Selection of Localization Radius in Ensemble Filtering for Multiscale Quasigeostrophic Dynamics. *Monthly Weather Review*, **146**, 543–560, https://doi.org/10.1175/MWR-D-17-0336.1.

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. *Monthly Weather Review*, **137**, 2105–2125, https://doi.org/10.1175/2009MWR2645.1.

Zhao, Q., and F. H. Carr, 1997: A Prognostic Cloud Scheme for Operational NWP Models. *Monthly Weather Review*, **125**, 1931–1953, https://doi.org/10.1175/1520-0493(1997)125<1931:APCSFO>2.0.CO;2.

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. *Bulletin of the American Meteorological Society*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Tables

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 ignored in the estimation.

	Experiment	SR-High	MR170	MR204
4DEnVar	# of Cores	1260	1260	1260
	Time (h)	0.92	0.91	0.91
	Core Hours	**1159.20**	**1146.60**	**1146.60**
LETKF	# of Cores	1260	1260	1260
	Time (h)	0.21	0.67	0.90
	Core Hours	**264.60**	**844.20**	**1134.00**
Control Member Forecast	# of Cores	396	396	396
	Time (h)	0.079	0.079	0.079
	Core Hours	**31.28**	**31.28**	**31.28**
High-Res Ensemble Forecast	# 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-Res Ensemble Forecast	# 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

Figures

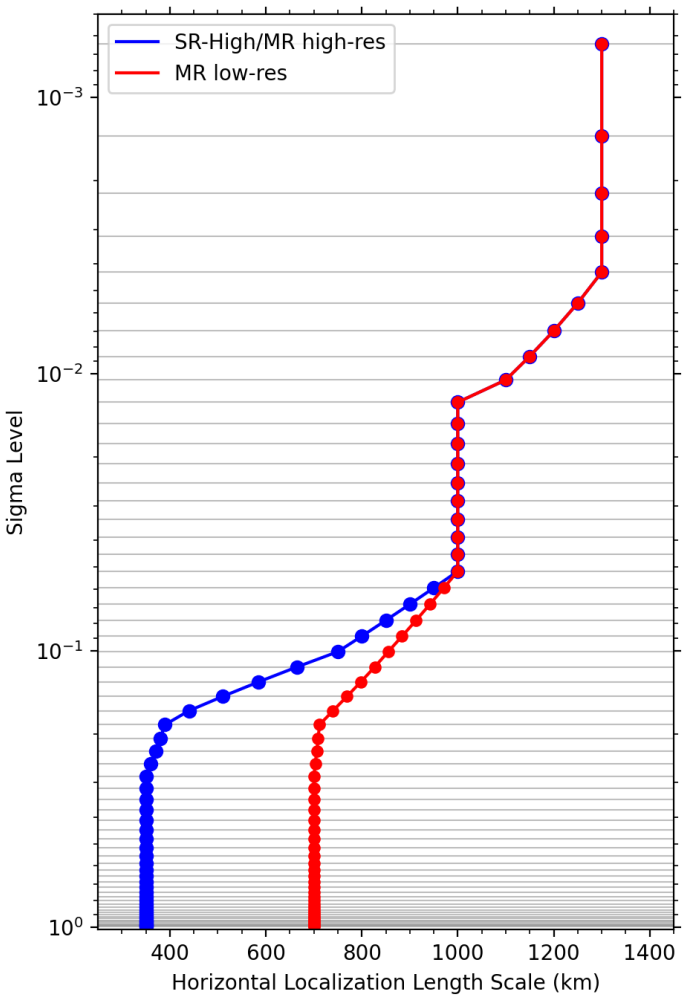

Fig. 1. Level-dependent localization length scales in e-folding distance (km) for 4DEnVar for the SR-High experiment and high-resolution ensemble for the MR experiments (blue) and for the low-resolution ensemble for the MR experiments (red).

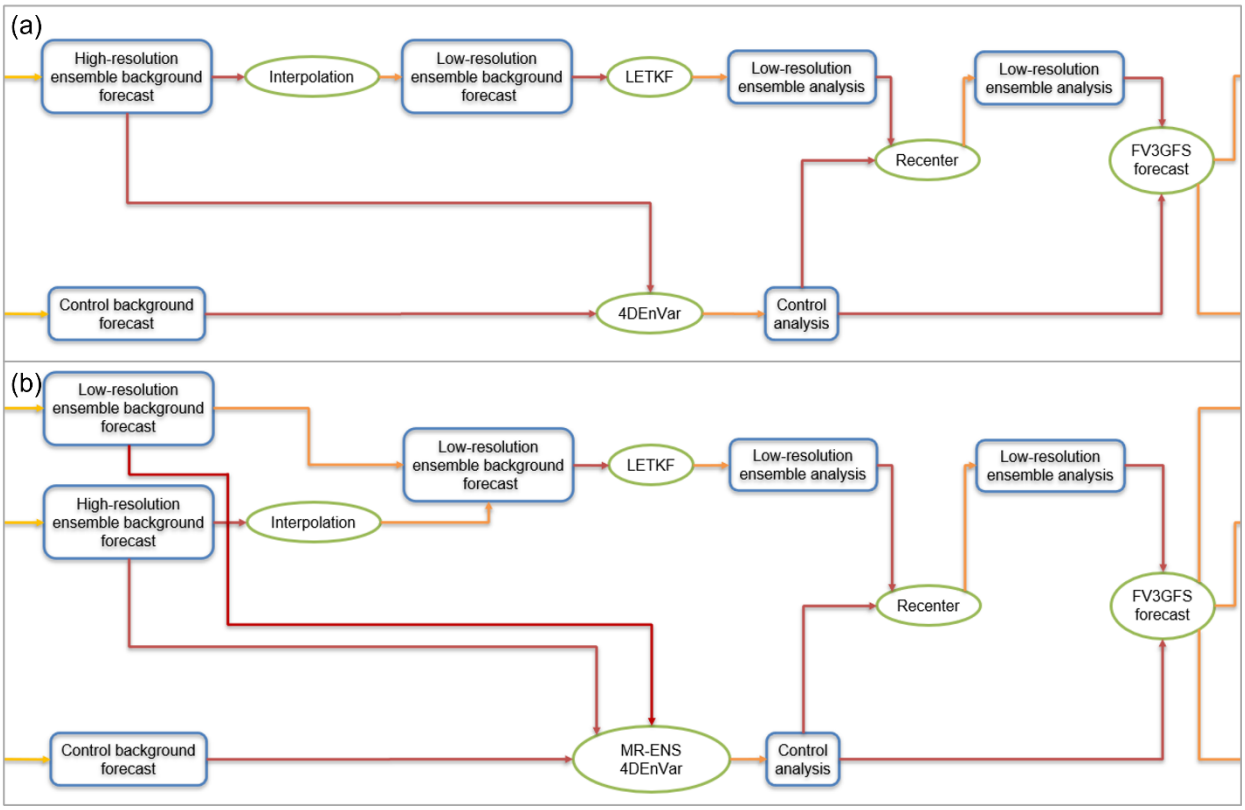

Fig. 2. Flow charts for a cycle of the (a) SR-High and (b) MR experiments.

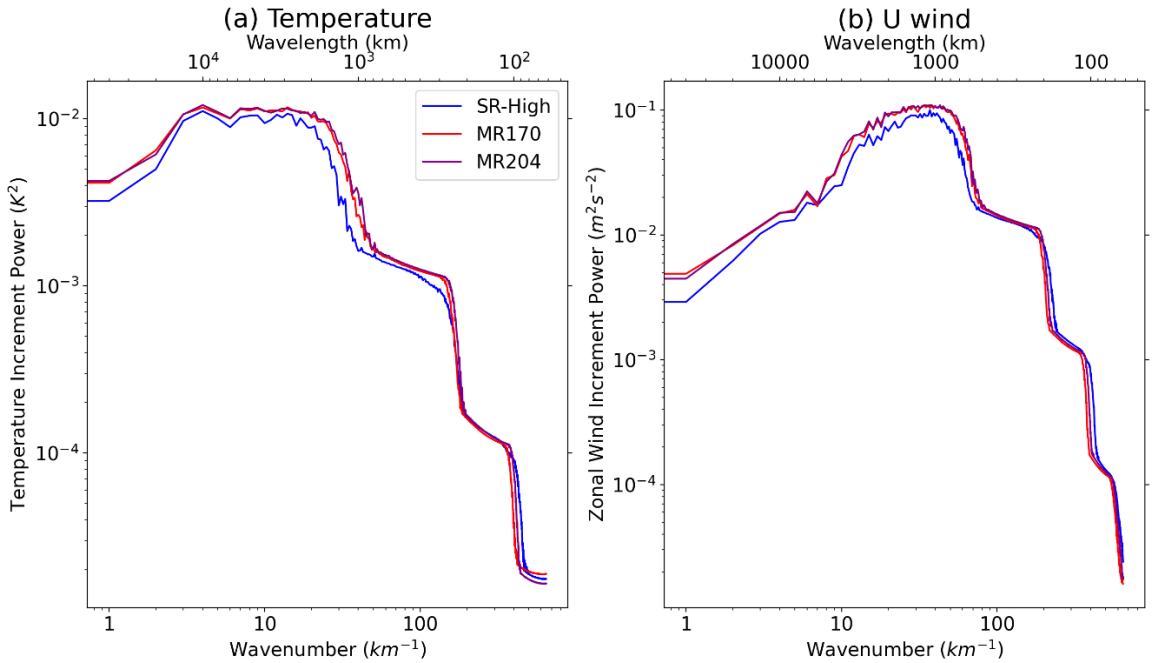

Fig. 3. Averaged analysis increment power at 500 hPa over all DA cycles as a function of wavenumber and wavelength for (a) temperature (K^2) and (b) zonal wind speed (m^2 s^{-2}) for the SR-High (blue), MR170 (red), and MR204 (purple) experiments.

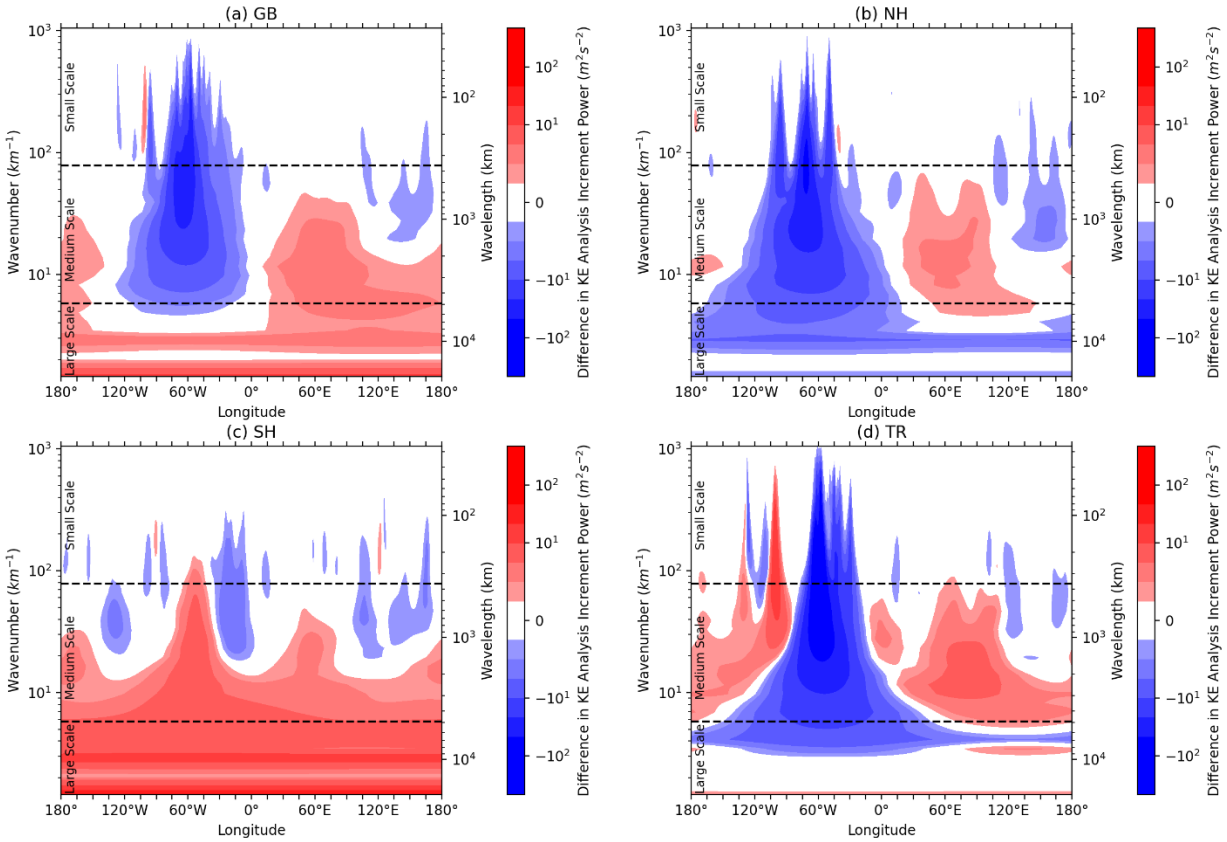

Fig. 4. Difference in kinetic energy wavelet analysis between MR170 and SR-High averaged over all DA cycles and over the (a) global, (b) Northern Hemisphere extratropical, (c) Southern Hemisphere extratropical, and (d) tropical latitudes. The x-axis represents the longitude. The left y-axis represents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

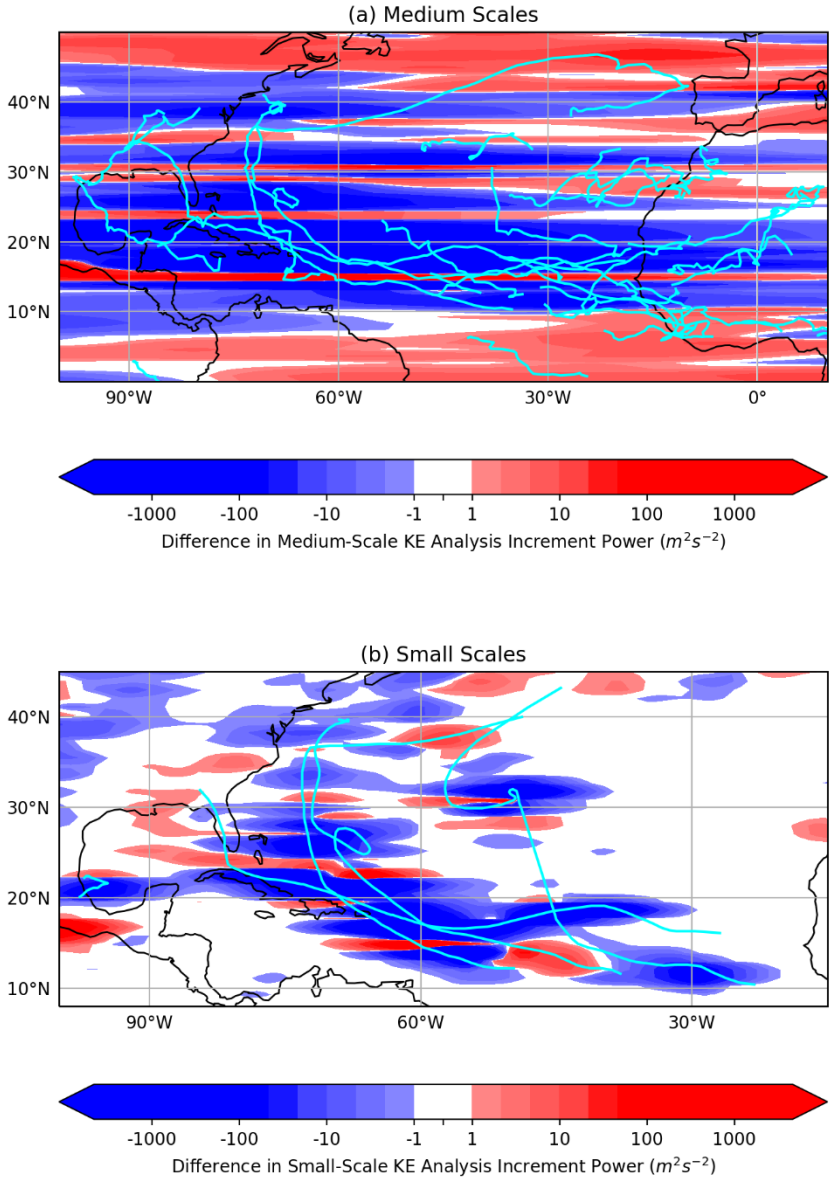

Fig. 5. Difference in the wavelet kinetic energy analysis increment power (m^2 s^{-2}) between MR170 and SR-High averaged over all cycles and over (a) medium scales (wavelet scales 16-30) and (b) small scales (wavelet scales 1-15). The tracks of (a) the tropical easterly waves that occurred at 850 hPa or (b) the Atlantic hurricanes that occurred during September 2017 are plotted in cyan.

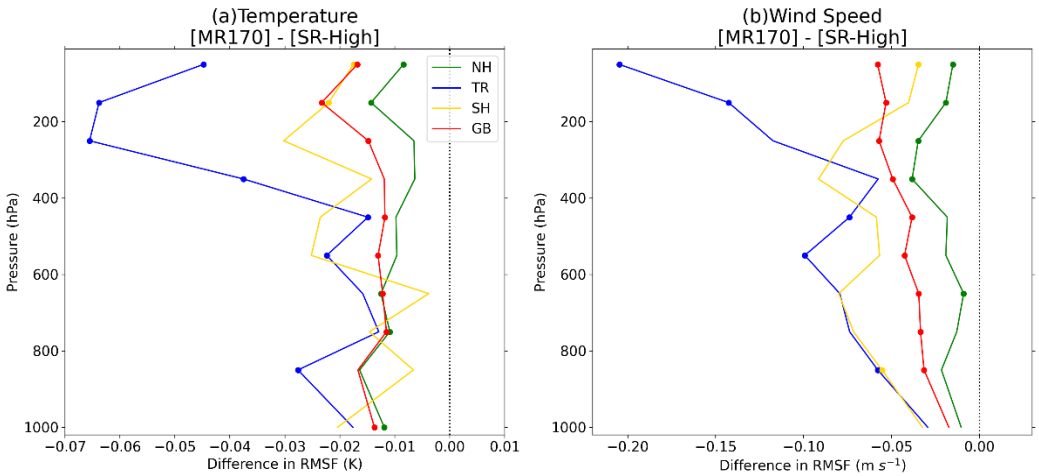

Fig. 6. Difference in the root mean square fit (RMSF) of the 6-h forecast of (a) temperature (K) and (b) wind speed (m s^{-1}) to the rawinsonde observations between the MR170 and SR-High experiments as a function of pressure, averaged over all cycles every 100 hPa over global (red), Northern Hemisphere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

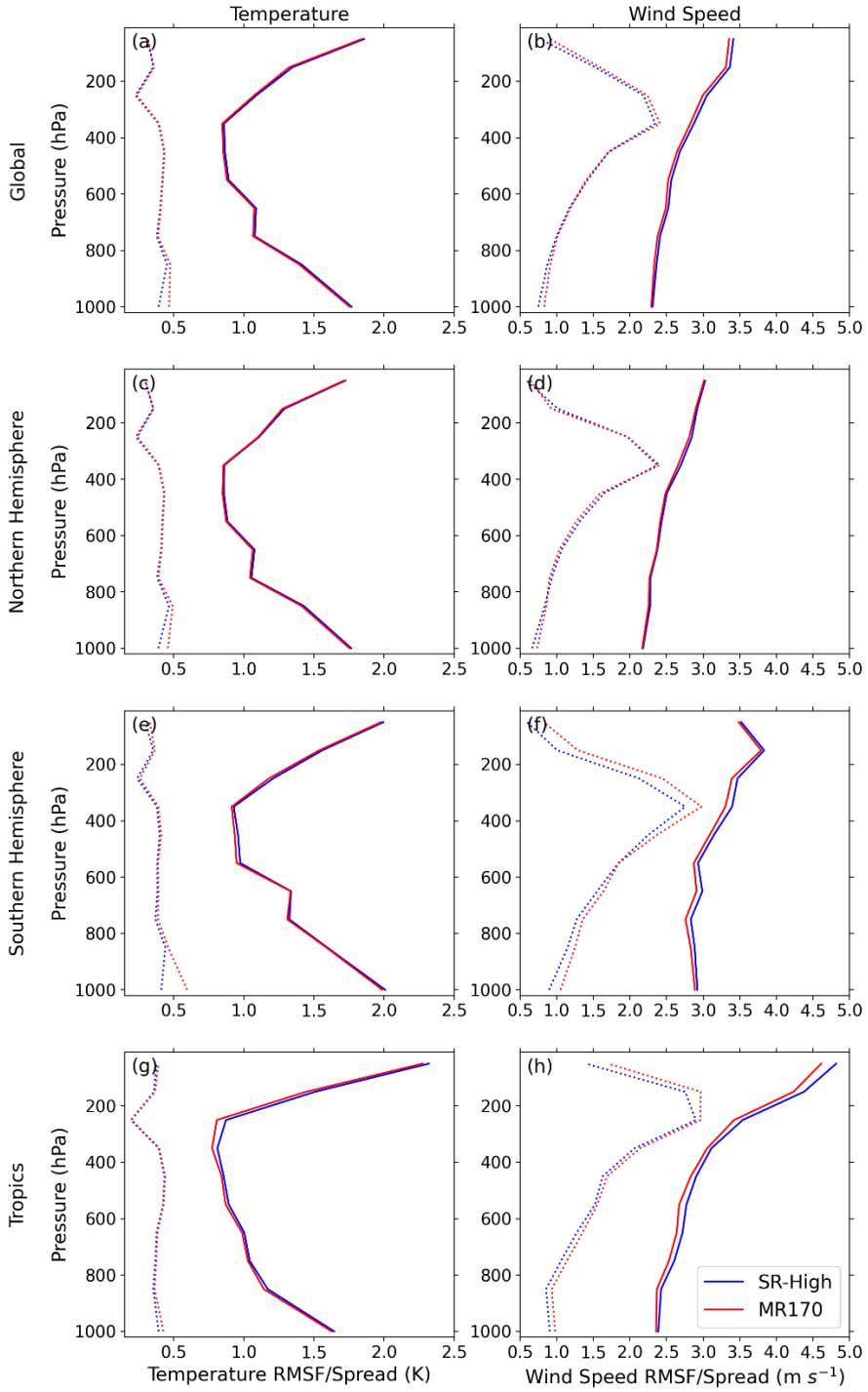

Fig. 7. The root mean square fit (RMSF) of the 6-h forecast to the rawinsonde observations (solid lines) and square root of the observation error variance plus the 6-h background ensemble variance (dotted lines) for the temperature (left; K) and wind speed (right; m s^{-1}) for the MR170 (red) and SR-High (blue) experiments as a function of pressure averaged over all cycles every 100 hPa over global (a,b), Northern Hemisphere extratropical (c,d), Southern Hemisphere extratropical (e,f), and tropical (g,h) latitudes.

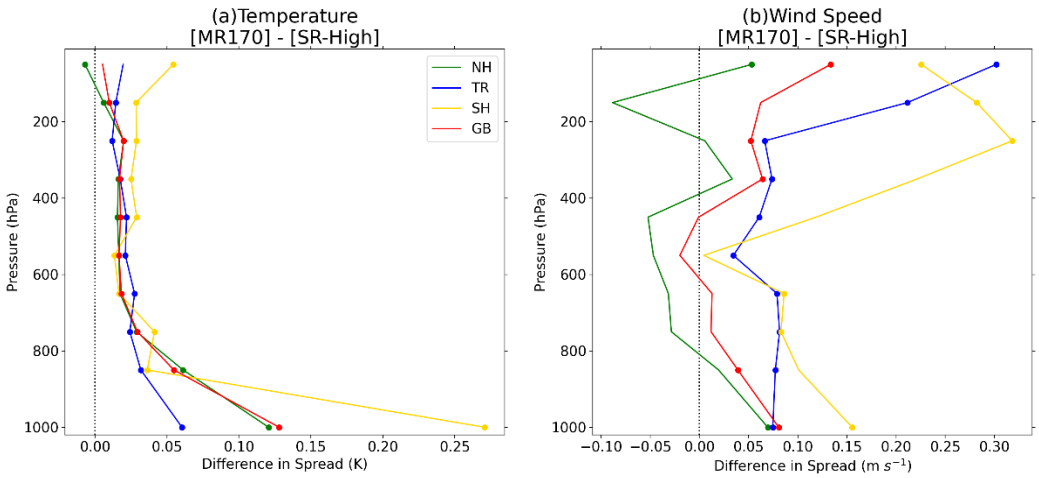

Fig. 8. As in Fig. 6 except of difference of the square root of the 6-h forecast ensemble variance (i.e., the ensemble spread) at locations of the rawinsonde observations.

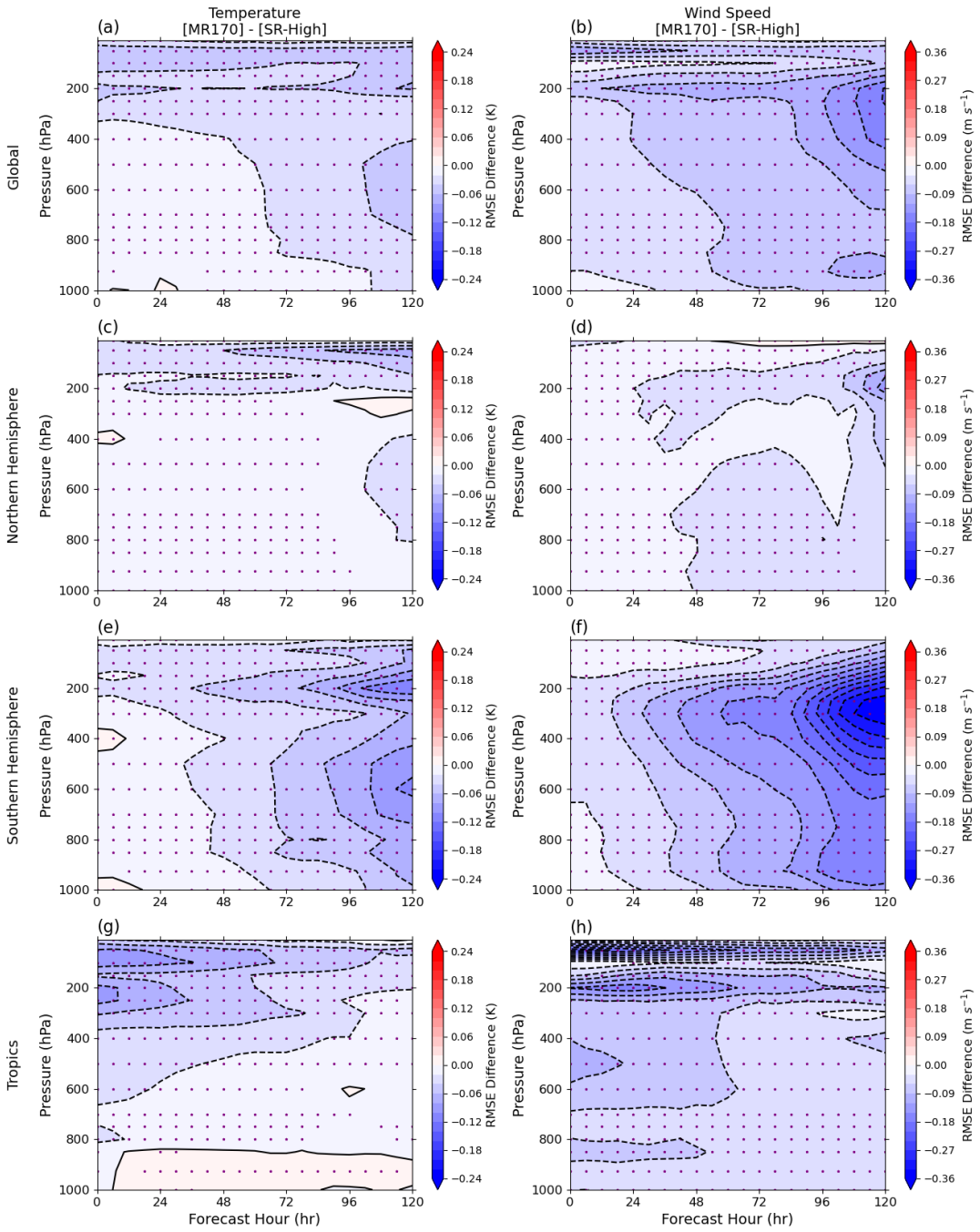

Fig. 9. Difference in root mean square error (RMSE) of the MR170 and SR-High forecasts verified against ERA-Interim temperatures (left; K) and wind speeds (right; m s^{-1}) averaged over all cycles for global (a,b), Northern Hemisphere extratropical (c,d), Southern Hemisphere extratropical (e,f), and tropical (g,h) latitudes for 5-days in lead time. Purple dots indicate levels where the RMSE difference is statistically significant using a paired-*t* test and a 95% confidence interval.

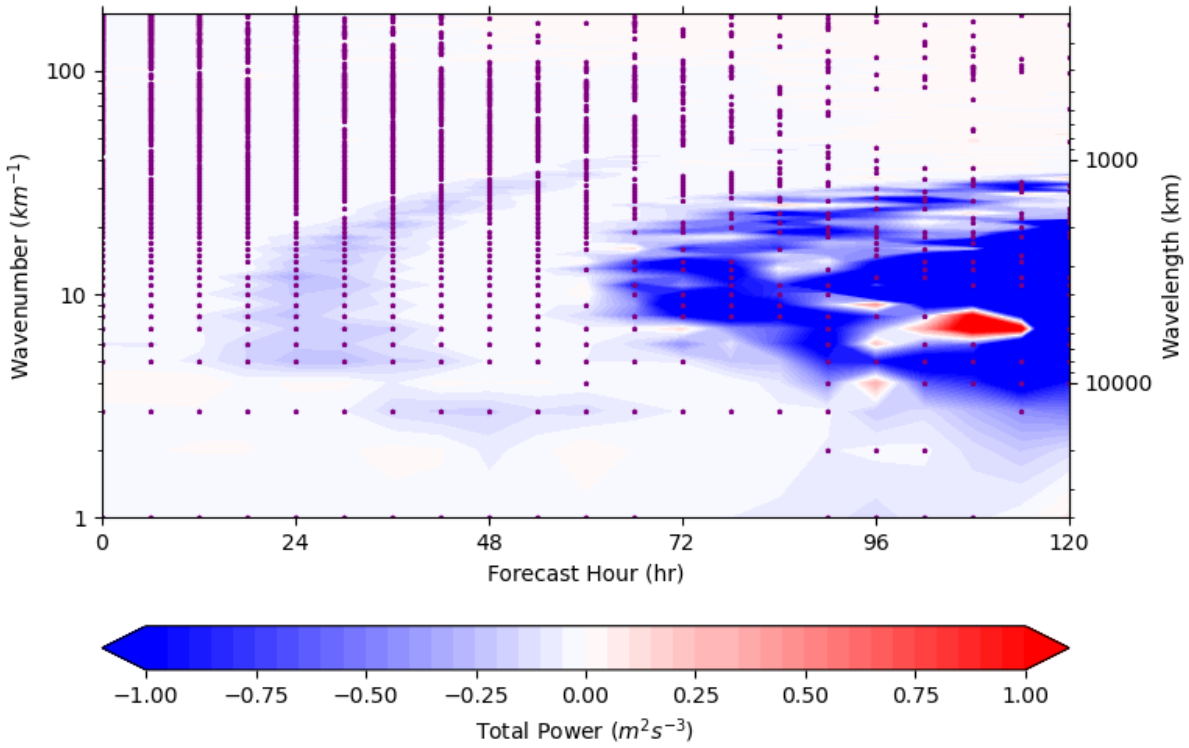

Fig. 10. Difference in total power error ($m^2 s^{-3}$) between the MR170 and SR-High experiments verified against ERA-Interim, averaged over all DA cycles and pressure levels selected in Fig. 9 and decomposed into wavenumber (left y-axis) and corresponding wavelength (right y-axis) for up to 5 days in lead time. Purple dots indicate levels where the difference is statistically significant using a paired-*t* test and a 95% confidence interval.

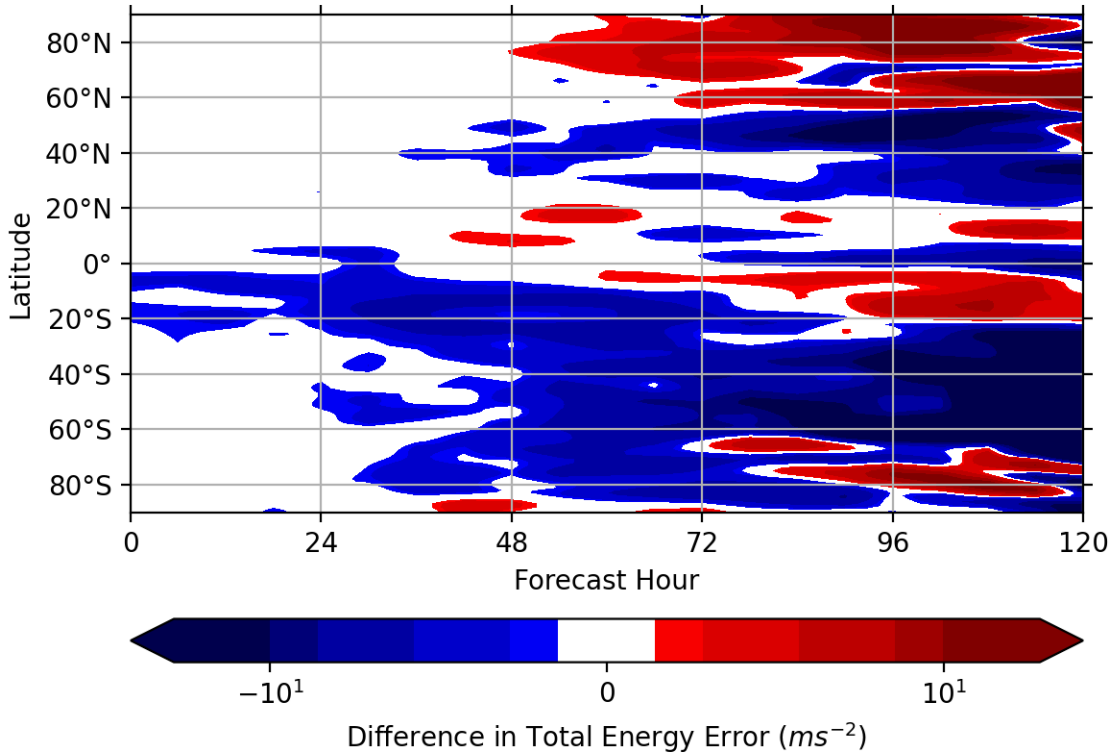

Fig. 11. Difference in total energy error ($m\,s^{-2}$) at 250 hPa filtered to include wavenumbers 5 to 25 for the 0000 UTC on 12 September 2017 cycle between the MR170 and SR-High experiments verified against ERA-Interim and averaged over all longitudes at a given latitude and up to 5 days in lead time.

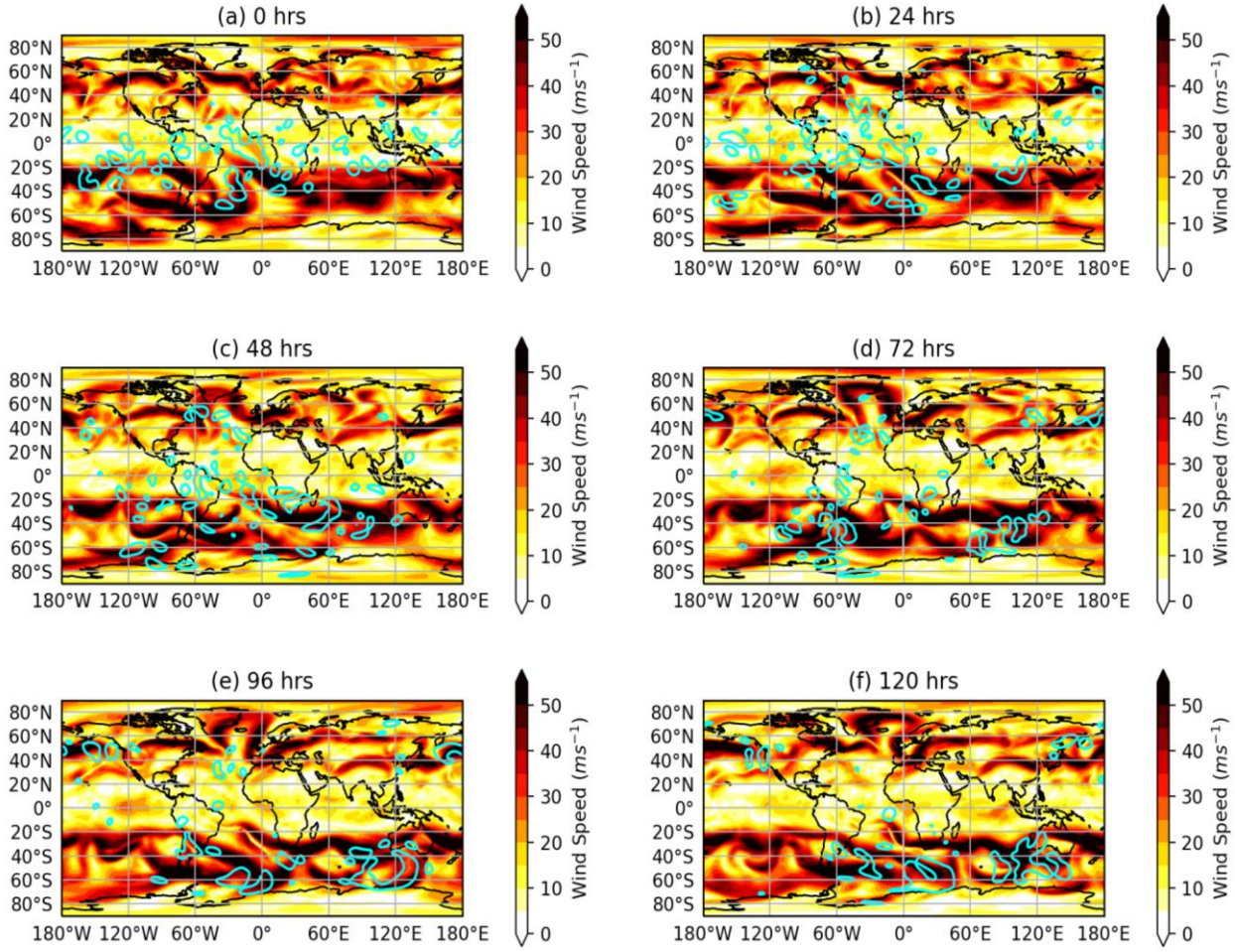

Fig. 12. ERA-Interim wind speed at 250 hPa for the 12 September 2017 at 0000 UTC cycle for the analysis time and every 24-h forecast up to 5 days. The 5% minimum difference in total energy error at 250 hPa relative to ERA-Interim between the MR170 and SR-High experiments filtered to include wavenumbers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

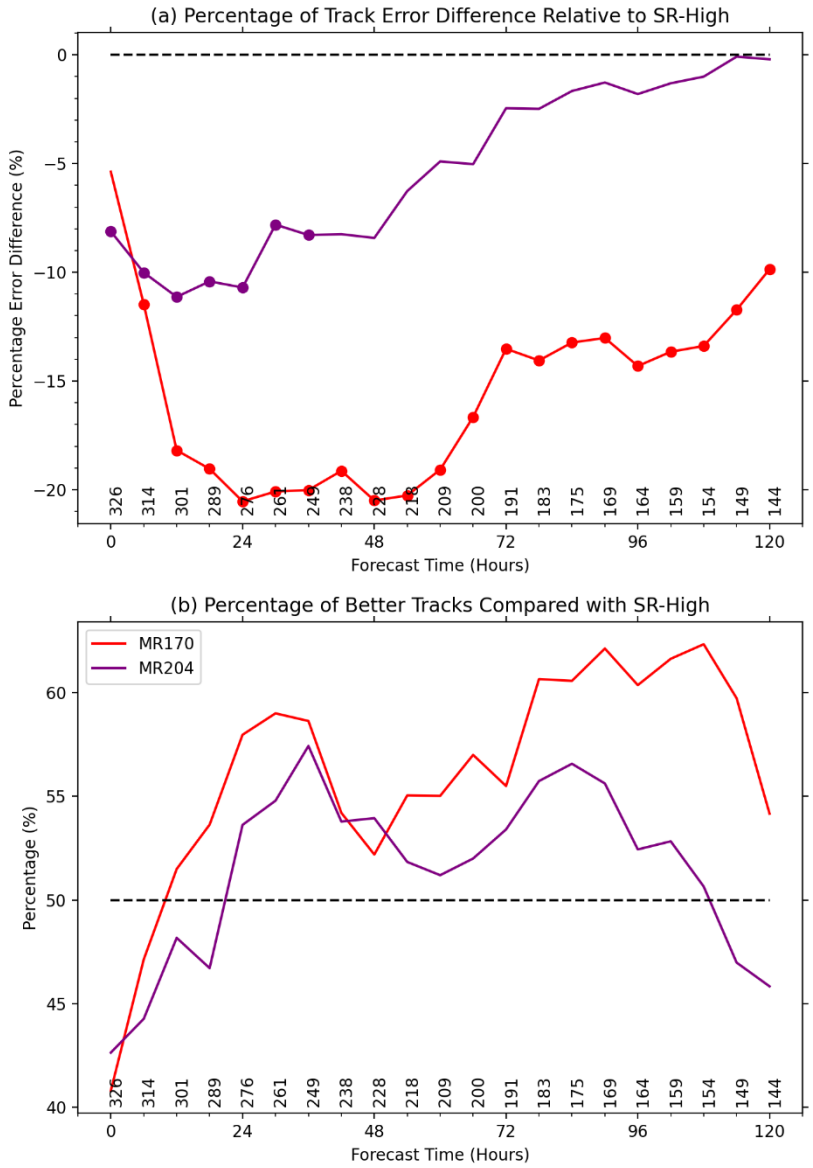

Fig. 13. (a) Percentage of track forecast error differences of MR170 (red) and MR204 (purple) relative to SR-High for lead times up to 5 days. Dots indicate levels where the difference is statistically significant using a paired-*t* test and a 95% confidence interval. (b) Percentage of track forecasts that are more accurate than SR-High with the same colors and lead times as in (a). The numbers above the x-axes indicate the sample size of tracks at the corresponding lead time.

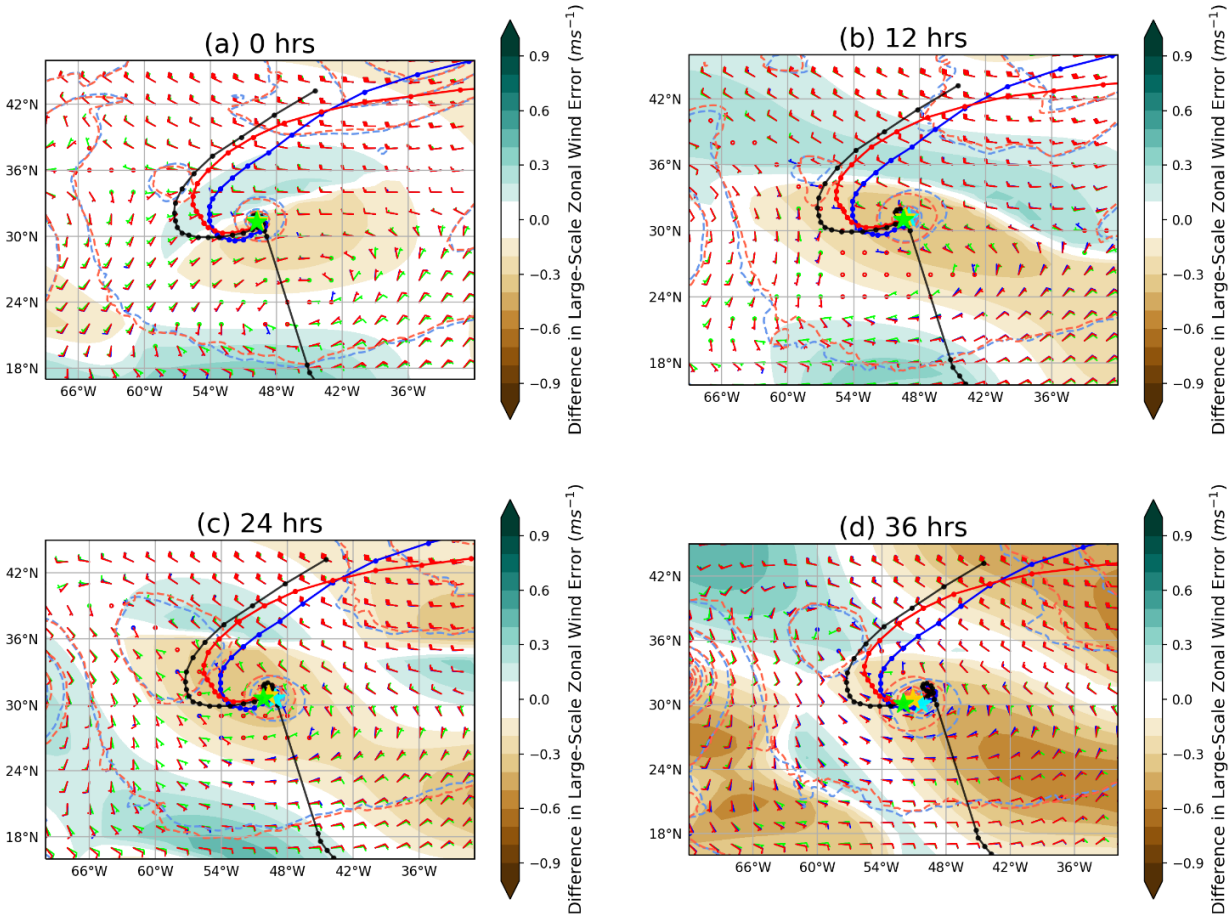

Fig. 14. Forecasted tracks of tropical cyclone Lee for MR170 (red solid), SR-High (blue solid) for lead times of (a) 0 h, (b) 12 h, (c) 24 h, and (d) 36 h with an initialization time of 1800 UTC on 24 September 2017, along with the best track of the storm (black solid). Dots are in 6-h increments for the tracks. Stars indicate the central position of the storm at the indicated lead times for the MR170 (gold), SR-High (cyan), and best (green) tracks. Mean sea level pressure is plotted every 4 hPa for the MR170 (red dashed) and SR-High (blue dashed) experiments. Wind barbs, indicating the large-scale (wavenumbers 0-10) pressure-weighted 850-300 hPa wind speed (kts) and direction, at are plotted for MR170 (red), SR-High (blue), and ERA-Interim (green) at each lead time. 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.

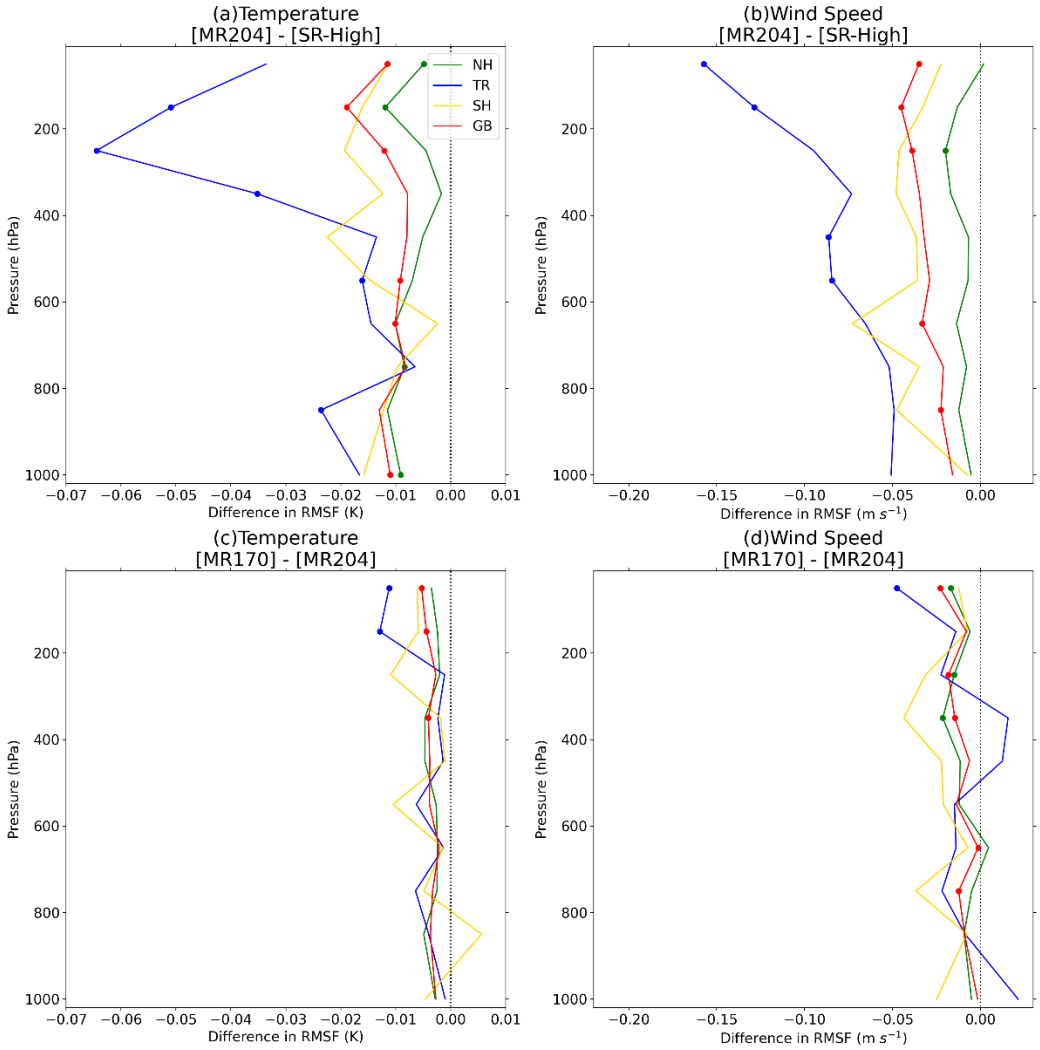

Fig. 15. As in Fig. 6 but (a,b) between the MR204 and SR-High experiments or (c,d) between the MR170 and MR204 experiments.

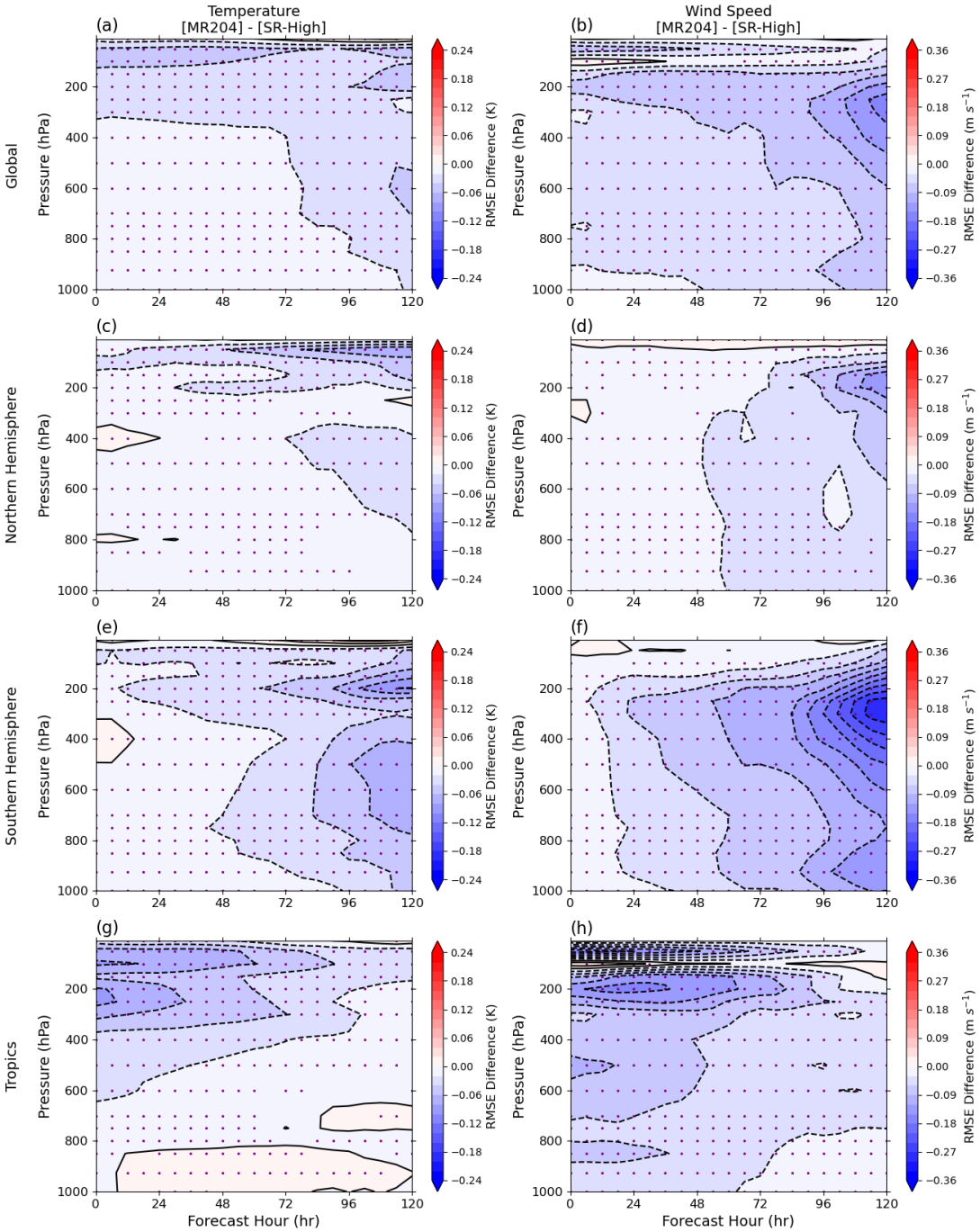

Fig. 16. As in Fig. 9 but for the MR204 and SR-High experiments.

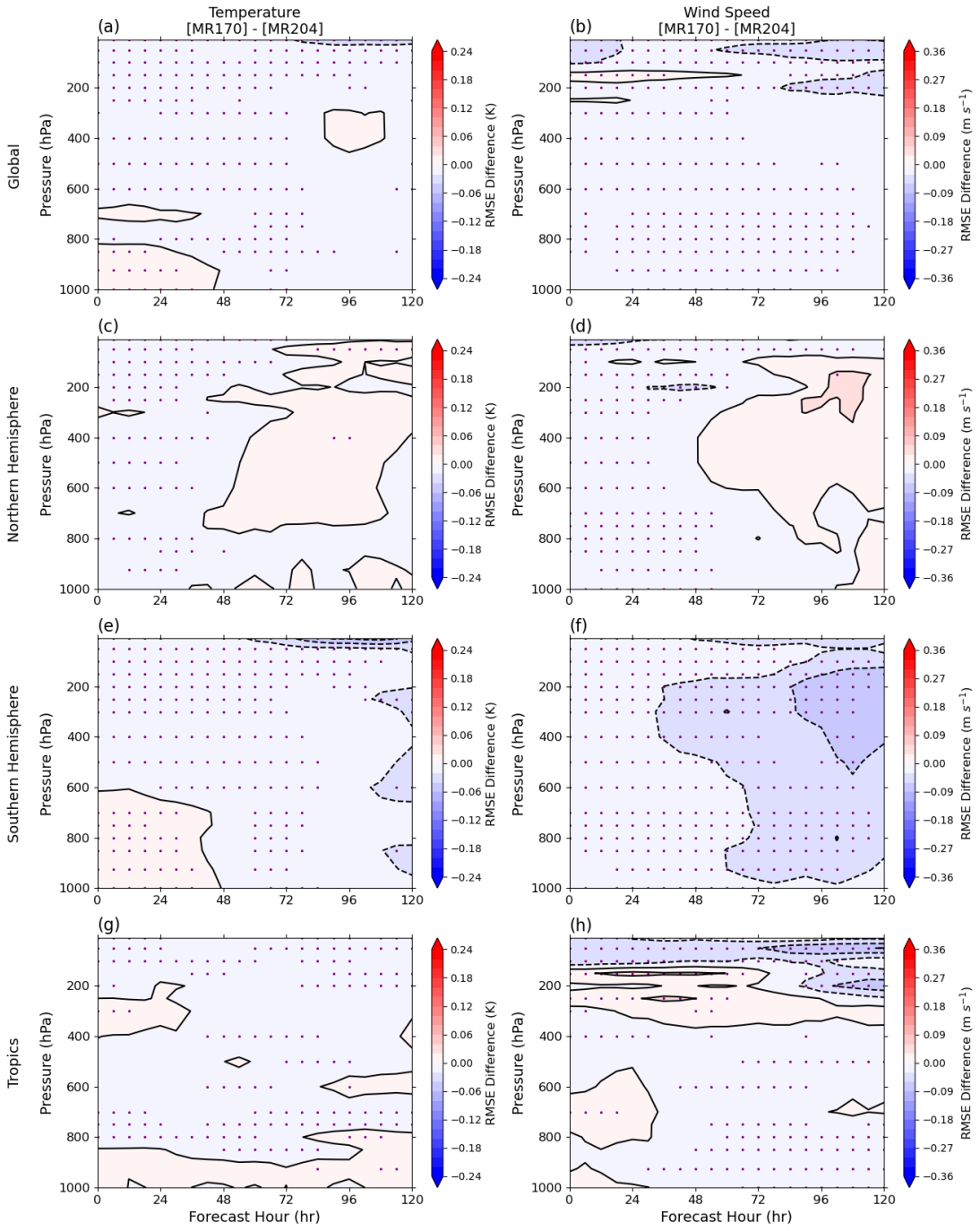

Fig. 17. As in Fig. 9 but for the MR170 and MR204 experiments.