OPERATIONAL PREDICTION SYSTEM NOTES

The Development of the NCEP Global Ensemble Forecast System Version 12

XIAQIONG ZHOU,[a] YUEJIAN ZHU,[b] DINGCHEN HOU,[b] BING FU,[c] WEI LI,[c] HONG GUAN,[d] ERIC SINSKY,[c]
WALTER KOLCZYNSKI,[c] XIANWU XUE,[d] YAN LUO,[c] JIAYI PENG,[c] BO YANG,[d] VIJAY TALLAPRAGADA,[b] AND
PHILIP PEGION[e]

[a] *CPAESS, University Corporation for Atmospheric Research at NCEP/EMC and GFDL, Princeton, New Jersey*
[b] *NOAA/NWS/NCEP/EMC, College Park, Maryland*
[c] *IMSG at NOAA/NWS/NCEP/EMC, College Park, Maryland*
[d] *SRG at NOAA/NWS/NCEP/EMC, College Park, Maryland*
[e] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*

(Manuscript received 9 July 2021, in final form 1 March 2022)

ABSTRACT: The Global Ensemble Forecast System (GEFS) is upgraded to version 12, in which the legacy Global Spectral Model (GSM) is replaced by a model with a new dynamical core—the Finite Volume Cubed-Sphere Dynamical Core (FV3). Extensive tests were performed to determine the optimal model and ensemble configuration. The new GEFS has cubed-sphere grids with a horizontal resolution of about 25 km and an increased ensemble size from 20 to 30. It extends the forecast length from 16 to 35 days to support subseasonal forecasts. The stochastic total tendency perturbation (STTP) scheme is replaced by two model uncertainty schemes: the stochastically perturbed physics tendencies (SPPT) scheme and stochastic kinetic energy backscatter (SKEB) scheme. Forecast verification is performed on a period of more than two years of retrospective runs. The results show that the upgraded GEFS outperforms the operational-at-the-time version by all measures included in the GEFS verification package. The new system has a better ensemble error–spread relationship, significantly improved skills in large-scale environment forecasts, precipitation probability forecasts over CONUS, tropical cyclone track and intensity forecasts, and significantly reduced 2-m temperature biases over North America. GEFSv12 was implemented on 23 September 2020.

KEYWORDS: Ensembles; Numerical weather prediction/forecasting; Operational forecasting

1. Introduction

Following the National Centers for Environmental Prediction (NCEP)'s implementation plan, a unified community model was developed acting as the foundation to align collaboration with the U.S. modeling community and build the NCEP Environmental Modeling Center (EMC) unified modeling capabilities. The GFDL Finite-Volume Cubed-Sphere (FV3) was chosen as the dynamical core for the Next Generation Global Prediction System (NGGPS) in 2016. The first major NGGPS model package was successfully implemented within the Global Forecast System (GFS) and became operational on 12 June 2019 as the GFSv15. The FV3-based model used in GFSv15 is the basis of the lower-resolution medium-range ensemble forecast system—the Global Ensemble Forecast System, version 12 (GEFSv12 hereafter).

One important strategy for the NCEP implementation is simplifying the NCEP production suite, as the production suite complexity and a large number of models are key factors that are limiting its effectiveness and ability to improve. The wave model (WAVEWATCH III; Tolman 2016) was integrated into FV3-based GEFSv12 with one-way coupling to replace the Global Wave Ensemble System. A second GEFS control member is added to run the GEFS-aerosol forecast to replace the current NEMS GFS Aerosol Component (NGAC; Wang et al. 2018). GEFSv12 with the integrated WAVEWATCH III and NGAC was implemented on 23 September 2020, as one operational system after extensive evaluation.

The Global Spectral Model has been used in GEFS since its first implementation in 1992 (Toth and Kalnay 1993). The GEFS version 11 (GEFSv11, hereafter) was implemented on 2 December 2015 (Zhou et al. 2017), and it uses a semi-Lagrangian Global Spectral Model (GSMv12.0.0). A tremendous amount of testing and evaluation was performed to build GEFSv12 to accomplish the big move from a legacy global spectral model with a hydrostatic assumption to a nonhydrostatic FV3-based model.

An experimental extended forecast system of GEFS was developed in 2018 with the support of the Subseasonal Experiment (SubX) project (Pegion et al. 2019). This system is based on the GEFSv11 configuration with upgrades to the model stochastic scheme and a minor upgrade to the convection scheme (Zhu et al. 2017, 2018; Li et al. 2018). It runs experimentally and provides once-a-week real-time forecasts to support the Climate Prediction Center (CPC) week-3/4 outlooks. This experimental system serves as a benchmark for the extended forecasts of GEFSv12.

GEFSv12 officially extends the weather forecast guidance from 16 to 35 days, which allows the National Weather Service (NWS) to deliver numerical weather predictions 3–4 weeks in

Corresponding author: Xiaqiong Zhou, xiaqiong.zhou@noaa.gov

DOI: 10.1175/WAF-D-21-0112.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

TABLE 1. The upgrade summary of GEFSv12 compared to GEFSv11.

	GEFSv11	SubX	GEFSv12
Model	GSM (hydrostatic)	GSM (hydrostatic)	FV3-based (nonhydrostatic)
IC uncertainty	EnKF with TC relocation	EnKF with TC relocation	EnKF without TC relocation
Model uncertainty	STTP	SPPT + SKEB + SHUM	SPPT + SKEB
Resolution	TL574L64 (~34 km), 0–8 days TL382L64 (~52 km), 8–16 days	TL574L64 (~34 km), 0–8 days TL382L64 (~52 km), 8–35 days	C384L64 (~25 km)
Forecast days	16 days	35 days	16 days (0600, 1200, and 1800 UTC) 35 days (0000 UTC)
Ensemble size	20 members	20 members	30 members
Ocean forcing	Persistent + relaxation SST	NSST and two-tiered SST	NSST and two-tiered SST
Microphysics	Zhao–Carr MP	Zhao–Carr MP	GFDL MP

advance. For the extended forecast system, reforecasts with a long training dataset are highly desirable to remove systematic errors in the prediction system (Hamill et al. 2006, 2008). The implementation of GEFSv12 is accompanied by the creation of a 20-yr global reanalysis dataset from 2000 to 2019 and a 30-yr reforecast dataset from 1989 to 2019, which were made available to the public (Hamill et al. 2021; Guan et al. 2022). The 20-yr reanalysis for the 2000–19 period was performed based on the GFSv15 configuration except for a lower horizontal resolution (Hamill et al. 2021). The 30-yr GEFS reforecasts are initialized from the 20-yr reanalysis data from 2000 to 2019 and the Climate Forecast System Reanalysis (CFSR; Saha et al. 2014) from 1989 to 1999. The reforecast uses the same model configuration as GEFSv12 but differs in the ensemble size, forecast lead time and the verification period. The verification of the 30-yr GEFSv12 reforecast was performed against the benchmark of GEFS extended forecasts (GEFS SubX) and GEFSv10 reforecasts (Hamill et al. 2013), which can be found in another paper (Guan et al. 2022). Overall, GEFSv12 outperforms the GEFS SubX and GEFSv10 in terms of the 500-hPa geopotential height, tropical cyclone track, precipitation over CONUS, and MJO forecasts.

The focus of this study is to introduce the sensitivity tests that we performed to determine the configuration of GEFSv12 and its performance compared with the operational-at-the-time version (GEFSv11) in more than 2-yr retrospective runs. The verification metrics to evaluate the GEFSv12 are introduced in section 2. The major upgrades in terms of the model dynamics, physics parameterization, and the ensemble configuration of GEFSv12 are introduced in section 3. The sensitivity experiments performed during the tuning and testing processes are discussed in section 4. The overall performance of the 2-yr retrospective runs is discussed in section 5. Section 6 is the summary and discussion.

2. Forecast verification metrics

The performance of ensemble forecasts is evaluated by using the NCEP ensemble verification system and verified against the corresponding analysis (Toth et al. 2003, 2006; Zhu and Toth 2008; Zhou et al. 2016, 2017), after the forecast and analysis fields are interpolated to a 2.5° × 2.5° latitude–longitude grid. The verification metrics for general forecast variables include the forecast bias and ensemble spread, the

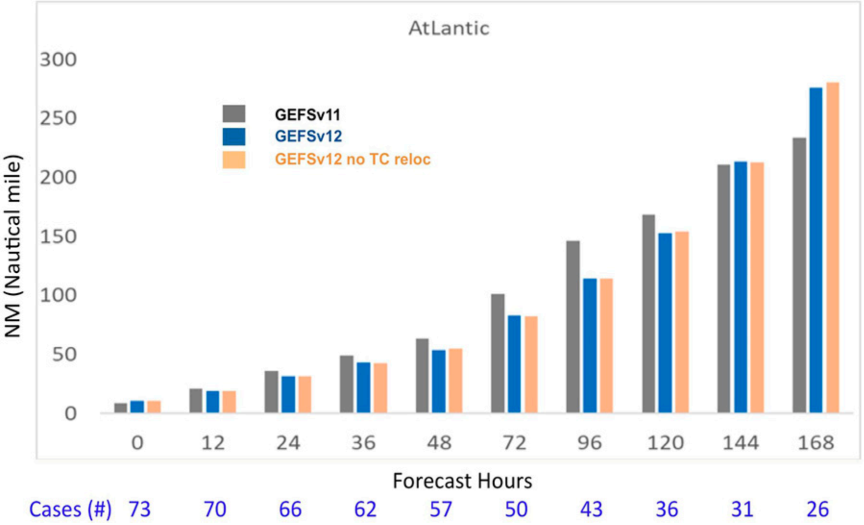

FIG. 1. Tropical cyclone track forecast errors for GEFSv11 (gray) and FV3-based GEFSv12 without (blue) and with (orange) tropical cyclone relocation over the Atlantic basin for the period from 16 Aug 2017 to 30 Sep 2017.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

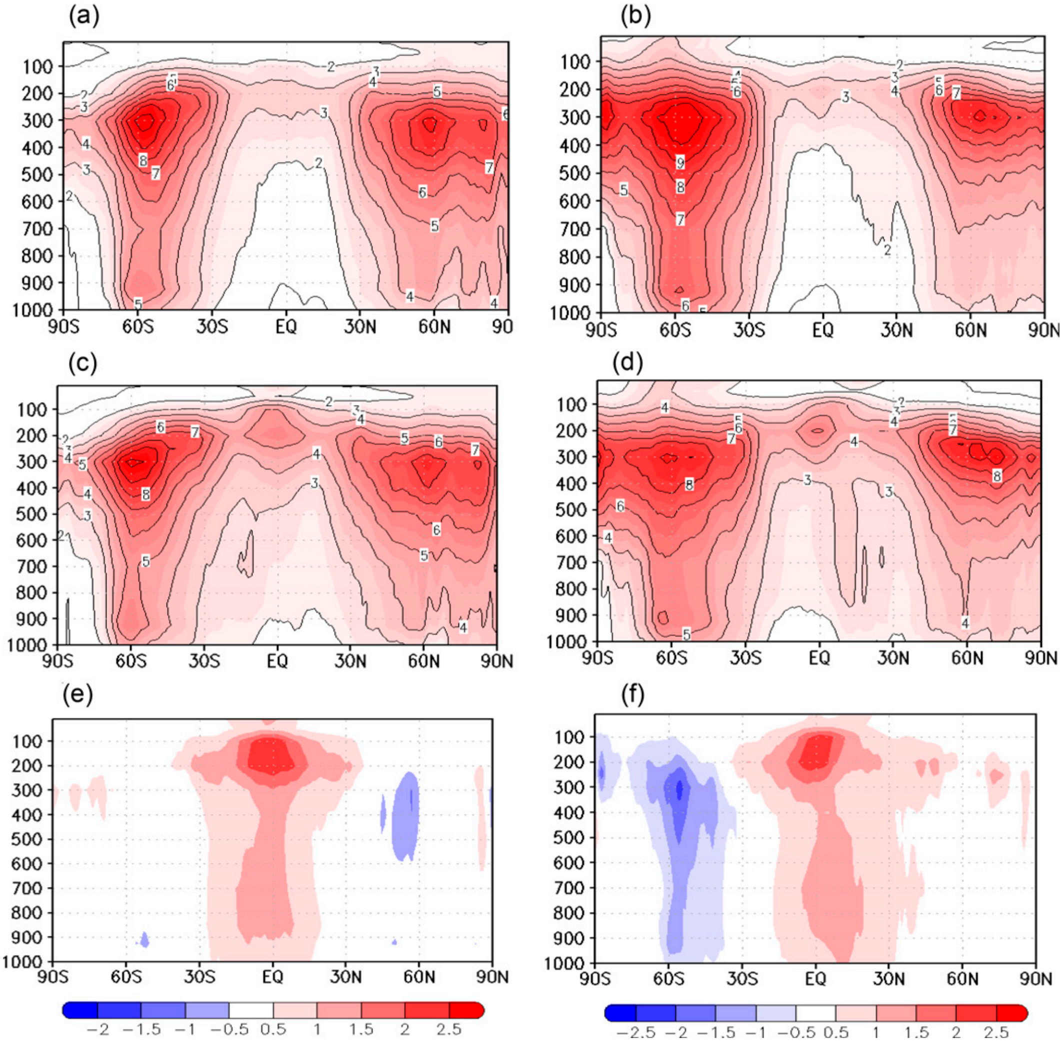

FIG. 2. The zonal-mean ensemble spread of zonal winds in (a) GEFSv11, (c) GEFSv12, and (e) the spread difference between these two versions in the Northern Hemisphere winter season. (b),(d),(f) As in (a), (c), and (e), but for the Northern Hemisphere summer season.

root-mean-square error (RMSE) and the pattern anomaly correlation (PAC) for the ensemble mean, and also probability forecast skill scores such as the continuous ranked probability score (CRPS), the continuous ranked probability skill score (CRPSS), the Brier score (BS), the Brier skill score (BSS) and the relative operating characteristic (ROC).

The verified fields include the geopotential height at 500 and 1000 hPa; wind fields at 10 m, 850 hPa, and 250 hPa; and temperature at 2 m and 850 hPa in the Northern Hemisphere (NH), the Southern Hemisphere (SH), and tropics. Our discussion mainly focuses on the 500-hPa geopotential height field as it is the standard means for evaluating the large-scale prediction skill of medium-range prediction models, except for when special concerns or attention to other forecast variables are needed. A paired block bootstrap method is used for the statistically significant test of the differences between two forecasts (Hamill 1999). The set of two forecasts is repeatedly randomly sampled to build the null distribution from which 2.5th and 97.5th percentiles are assessed to get the confidence interval.

The verification includes the quantitative precipitation forecast (QPF) over the contiguous United States (CONUS), tropical cyclone track and intensity forecasts, and MJO prediction skills. QPF and probabilistic QPF are verified against the climatology-calibrated precipitation analysis (CCPA) over CONUS. CCPA is generated with linear regression and spatial and temporal downscaling techniques and combining two widely used datasets: the NCEP CPC Unified Global Daily Gauge Analysis and the higher temporal and spatial resolution of the NCEP Stage IV multisensor quantitative precipitation estimations (Hou et al. 2014). Precipitation is categorized by the 24-h accumulated value with threshold amounts greater than 1, 5, 10, and 20 mm. BSS indicates the degree of improvement of BS of a forecast compared to the

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

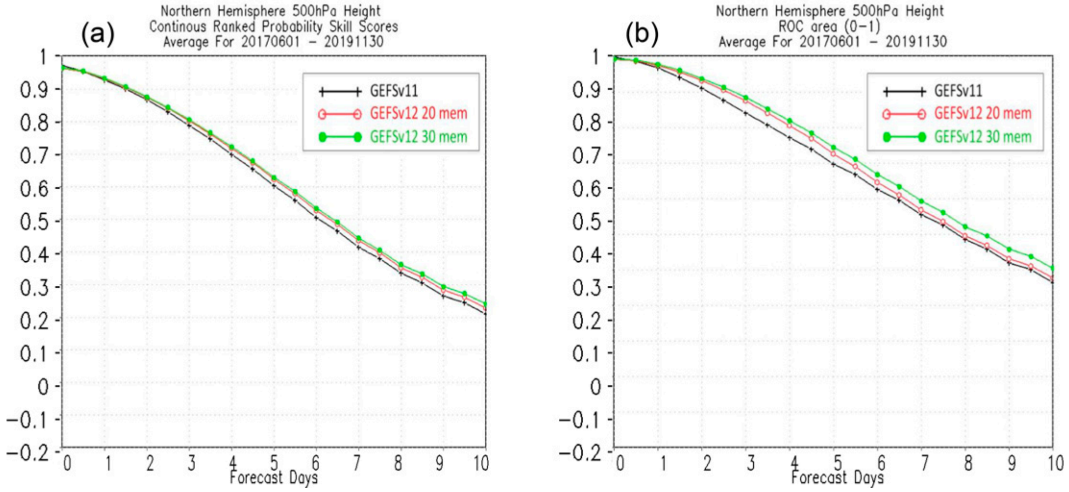

FIG. 3. The (a) CRPSS and (b) ROC for 500-hPa geopotential height in the Northern Hemisphere. The curves with different colors represent GEFSv11 (black) and GEFSv12 with 20 (red) and 30 (green) ensemble members.

climatology. The 10-yr mean of CCPA is used as the climatology to calibrate the BS. As the GEFSv12 extends the ensemble forecasts to a subseasonal time scale (35 days), the forecast skill of MJO as the main prediction source at this time scale is also compared with the GEFS SubX.

3. Major upgrades of GEFSv12

GSM in NCEP had been developed and used in GFS and GEFS for over 30 years, but it remained a hydrostatic model without having nonhydrostatic extension. FV3 was selected from five nonhydrostatic candidate dynamical cores (Ji and Toepfer 2016) to meet the end goal of the NGGPS program for developing a single integrated model suitable for both global and regional modeling with forecast length ranging from weather to climate scales. GEFSv12 is the first FV3-based version of the GEFS system. The GFDL FV3 team has a detailed description of FV3 in their published papers (Lin and Rood 1996, 1997; Lin 1997, 2004; Harris and Lin 2013; Putman and Lin 2007; Harris et al. 2020a,b).

GEFSv11 became operational in 2015 with a horizontal resolution of T_L574 (34 km) for the first eight days and T_L384 (52 km) for the second eight days. There are 64 vertical levels on sigma pressure hybrid layers. The initial conditions for the 20 ensemble members are generated from a GSI/EnKF hybrid analysis perturbed with 6-h EnKF ensemble forecasts (Zhou et al. 2016). The stochastic total tendency perturbation (STTP) scheme is used to represent model uncertainties by perturbing the total tendency of the model prognostic variables (surface pressure, temperature, wind, and humidity) with an empirical formula (Hou et al. 2006, 2008).

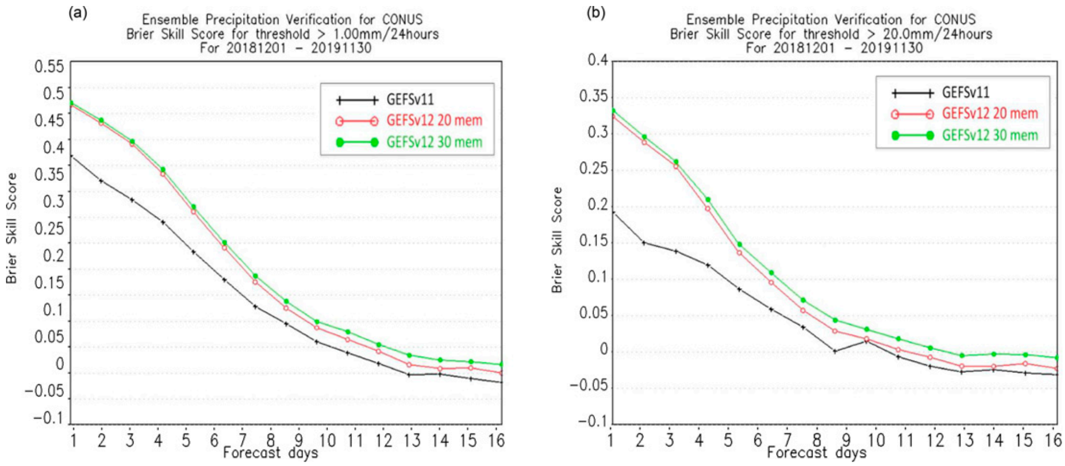

FIG. 4. The BSS for precipitation probability forecasts over CONUS with the precipitation threshold of (a) >1 and (b) >20 mm day^{-1}. The curves with different colors represent GEFSv11 (black) and GEFS v12 with 20 (red) and 30 (green) ensemble members.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

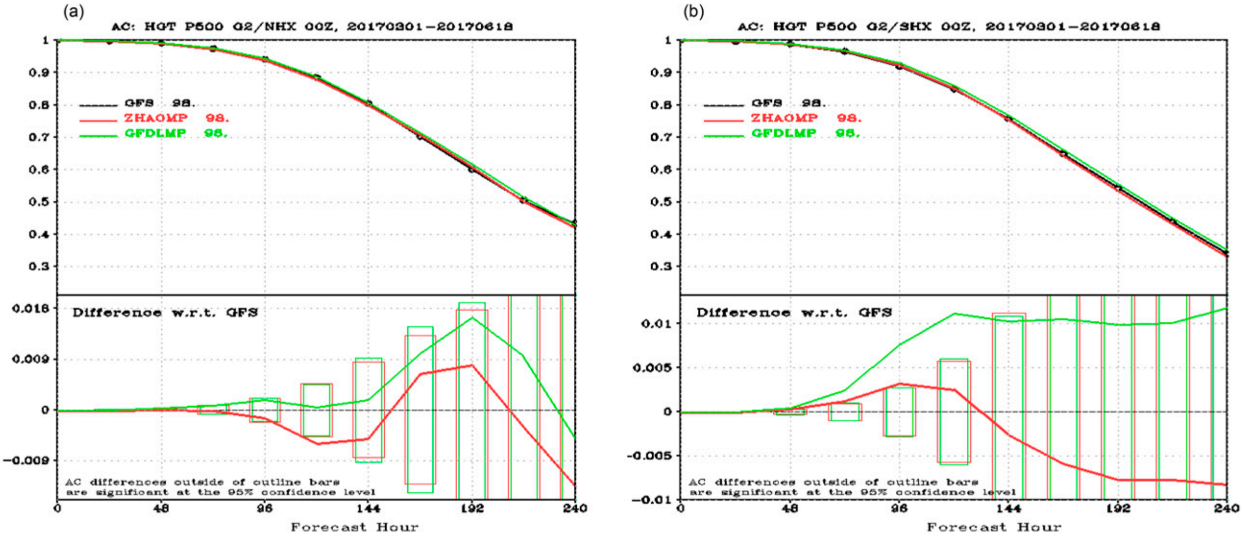

FIG. 5. The anomaly correlation for 500-hPa geopotential height in the (a) Northern Hemisphere and (b) Southern Hemisphere. The black curves represent the GFS, and the red and green curves represent the FV3-based GEFS with the Zhao–Carr MP and the GFDL MP. The lower graphs show the difference (curves) and significance test (bars). The difference is significant at the 95% confidence level when the value is outside the bars.

Table 1 lists the major upgrades of GEFSv12 from GEFSv11, as well as the GEFS SubX. The GEFS SubX as an experimental GEFS extended forecast system is similar to GEFSv11 except with an extended forecast length and new model stochastic schemes. The major GEFSv12 updates compared with its previous version include the replacement of the hydrostatic global spectral model (GSM v12) with the GFDL FV3-based nonhydrostatic model. The GFDL cloud microphysics scheme with five predicted cloud species (cloud water, cloud ice, rain, snow, and graupel; Zhou et al. 2019) replaces the Zhao–Carr microphysics scheme with only the total cloud water. In addition, the new GEFS was built on 25-km quasi-uniform grids having six tiles globally, with each tile having 384 × 384 grid cells. In contrast to the GEFSv11, which has two different horizontal resolutions with a higher resolution the first 8 days and a lower resolution the last 8 days, the GEFSv12 extended the forecast length from 16 to 35 days with a uniform horizontal resolution (about 25 km) through the entire model integration, and with an increase in ensemble members to 30 from the original 20. For the model uncertainty representation, the STTP scheme is replaced by a new stochastic physics suite.

The physics package in GEFSv12 remains similar to the one used in GFSv15. Compared to GEFSv11, there are some minor updates to the deep- and shallow-convection schemes, the land model, and the ozone photochemistry scheme, except for the replacement of the microphysics parameterization (MP) scheme. The simplified Arakawa–Schubert (SAS) shallow and deep convection schemes (Han and Pan 2011) were updated with a scale-aware parameterization (Han et al. 2017). The convection scheme is also modified to reduce excessive cloud-top cooling, to stabilize the model. Other physics updates also include revised bare-soil evaporation to reduce a dry and warm bias, an updated parameterization of ozone photochemistry with additional production and loss terms (McCormack et al. 2006), and a new parameterization of middle atmospheric water vapor photochemistry (McCormack et al. 2008).

A near-surface sea temperature (NSST) model is used to predict the vertical profile of sea temperature between the surface and a reference level (about 5 m) by only considering two physical processes: diurnal thermocline layer warming and thermal skin layer (also known as sublayer) cooling. The sea temperature at the reference level is also called the foundation temperature in NSST. It is determined by using a two-tiered method as in the SubX GEFS, in which the Real Time Global (RTG) SST analysis with climatology tendencies converges to the Climate Forecast System (CFS) bias-corrected predictive SST in a 35-days' time scale. This scheme allows the surface temperature over the ocean to have diurnal variability and provides a more realistic thermal boundary condition for the atmosphere. Previous studies have shown that the two-tiered scheme represents the variation of ocean temperature forcing better than that of the persistence method and can result in improved MJO forecast skills (Zhu et al. 2019; Li et al. 2018).

4. Sensitivity experiments

Similar to the development of other NCEP forecast systems, the process of GEFS development is iterative. The overall project is broken into many subcomponents that are developed in parallel, by which GEFS developers have a quickly developed version that is intentionally incomplete, instead of a final version. The subcomponents of GEFS include preprocessing to generate ensemble members, utilities required for model postprocessing, model dynamics and physics schemes, and model uncertainty schemes. A group of

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

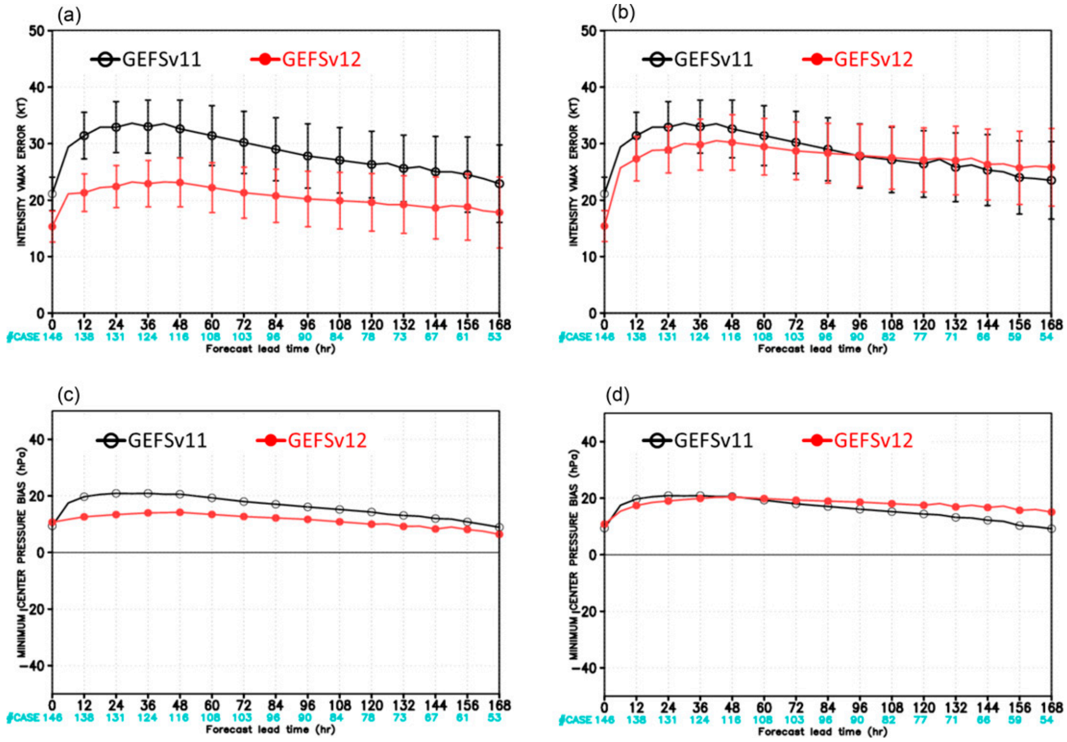

FIG. 6. Tropical cyclone intensity forecast error over the Atlantic basin with (a) HORD6 and (b) HORD5. (c),(d) As in (a) and (b), but for intensity bias. The blue curve represents GEFSv11, and the red is GEFSv12. The range of bars at (a) and (b) represents the 95% significance level.

sensitivity experiments is usually performed corresponding to each or several combined components. An appropriate mixed strategy is chosen based on the review of a series of sensitivity tests concerning program risks, performance, and computer resource usage. Finally, the integrated version is constructed based on the refined prototype. Two to three years of retrospective runs are performed and evaluated.

This development method means that the impact of each component on the GEFS performance cannot be demonstrated individually. The model used in the sensitivity experiments performed at the early stage is usually not the final version. In addition, sensitivity experiments only cover a short time period due to limited computing resources. Nevertheless, a clear comparison is made from sensitivity experiments with and without the potential upgrade components. The major sensitivity experiments that we performed in the testing period are summarized as the following.

a. *Initial condition uncertainties*

As in GEFSv11, the control run in GEFSv12 uses the hybrid GFS analysis as the initial conditions, while the ensemble members utilize the analysis perturbed by the 6-h EnKF forecast ensemble. In GEFSv11, tropical cyclones (TCs) in the ensemble members are separated from the environment by applying Kurihara's method (Kurihara et al. 1993, 1995). TC perturbations are calculated from the differences between ensemble members and the ensemble mean after the separated TC vortices are relocated to the observed locations (Liu et al. 2006; Zhou et al. 2017). The TC perturbations are rescaled based on tropical cyclone intensity and added to the analysis. The TC relocation process results in the tropical cyclone structure being perturbed, while the initial TC locations in the ensemble members are left intact. The impact of this TC relocation process is reevaluated for the GEFSv12. The sensitivity experiments, with and without this preprocessing, for one hurricane season (2017) show that there are no significant differences in terms of the ensemble mean track forecasts at all forecast lead times (Fig. 1). The track spread is slightly larger at the initial time without the TC relocation process, but it becomes almost identical after 24 h as the forecast times increase (not shown). The TC relocation process is not used in the new GEFS version due to the absence of any significant advantage. Note that the track forecast errors in GEFSv12 are generally smaller than those in GEFSv11, except for day 7. The degradation at day 7 is not considered a critical issue due to the small sample size. The exclusion of the TC relocation process in GEFSv12 simplifies the initialization process and reduces computational resources by about 10 min.

b. *Model uncertainties*

The STTP scheme was introduced in the GEFS in 2010. It is used in GEFS v9 to v11 to represent model uncertainty (Hou et al. 2006, 2008). This scheme adds stochastic forcing every 6 h to the total tendencies of the model prognostic variables. The stochastic forcing is calculated from the differences

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

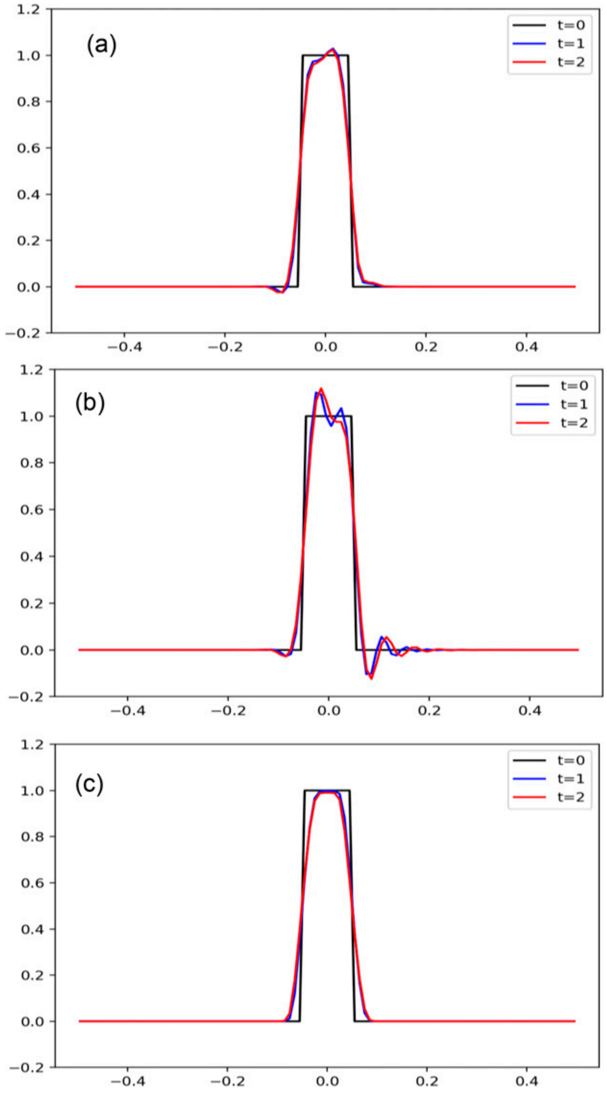

FIG. 7. The transport of a rectangular wave with periodic boundary conditions using the horizontal advection option (a) HORD5, (b) HORD6, and (c) HORD8. The black curve is for the initial time, while the blue and red curves represent the results after one and two periodic cycle times, respectively.

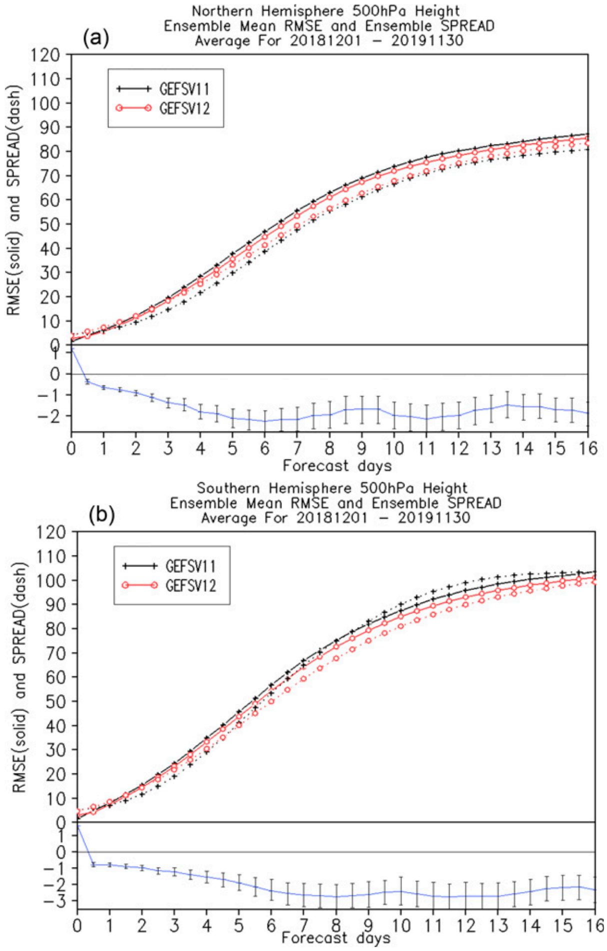

FIG. 8. The RMSE (solid lines) and ensemble spread (dotted) for 500-hPa geopotential height in the (a) NH and (b) SH. The black curves represent GEFSv11, and the red ones represent GEFSv12. The lower graphs show the RMSE difference and bootstrap significance tests. The difference is significant at the 95% confidence level when the bars do not overlap with the horizontal axis through zero.

in the total tendency changes between each ensemble member and the control after being multiplied by a random number and a rescaling factor as a function of location and lead time. Note that STTP requires all members to integrate simultaneously and communicate every 6 h, which could reduce the operational stability and reliability of GEFS.STTP is replaced by a new stochastic physics (SP) suite with two schemes in GEFSv12: stochastically perturbed physics tendencies (SPPTs; Buizza et al. 1999; Palmer 1997, 2001) and stochastic kinetic energy backscatter (SKEB; Berner et al. 2009; Shutts 2005). Unlike STTP, the new SP suite perturbs each ensemble member independently. Both SPPT and SKEB use a random pattern generator to generate spatially and temporally correlated random patterns in spectral space with first-order autoregressive [AR(1)] processes. These random patterns are then transformed to a Gaussian grid and interpolated to the model's native cubed-sphere grid. The SPPT perturbations applied to the physics tendencies are a sum of five different patterns with varying scales and amplitudes. The five horizontal length/time scales are 500 km/6 h, 1000 km/3 days, 2000 km/30 days, 2000 km/90 days, and 2000 km/1 year. Energy dissipation from resolved scales to subgrid scales in numerical weather prediction models occurs for numerical and physical reasons. SKEB aims to inject back a fraction of this to resolved scales. Kinetic energy dissipation from numerical diffusion and interpolation is added back in the FV3-based model. The kinetic energy loss of SKEB is calculated in the dynamic core as a heat source that could optionally be added to the temperature equation. SKEB uses this estimate after applying multiple passes of a Laplacian smoother. The horizontal length/time scales and the parameters that

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

control the perturbations pattern and amplitude of SKEB and STTP were carefully tuned to achieve a better ensemble spread–error relationship globally. In the early testing period, several crashes occurred, and the diagnostics indicated that the interaction of the PBL scheme and mountain blocking scheme with SPPT led to model instability. A fix is applied to SPPT with near-surface tampering below dividing streamlines, a parameter diagnosed in the gravity wave/mountain blocking scheme based on orography and kinetic energy (Lott and Miller 1997). This fix results in a slightly reduced ensemble spread of 2-m temperature (not shown) but significantly improved computational stability.

Early testing also indicates an increase in the global mean precipitation with SPPT. The bias comes from large increases in precipitation in areas of stratiform precipitation. Upon investigation, we discovered that the inconsistency between the moist physics tendencies that are perturbed and precipitation produced by the parameterizations, which are not perturbed, is the cause of the bias. The solution is to perturb the precipitation rate with the same random pattern that perturbs the physics tendencies. This removes the precipitation forecast bias and creates more physical consistency in the water budget.

The stochastically perturbed humidity (SHUM) is another component of the EMC stochastic physics suite. It is employed in the EnKF short-range forecast of the operational GFS GDAS. The SHUM scheme perturbs the PBL humidity, contributing to its importance in triggering convection in the physics parameterizations. The zonal mean ensemble spread from each SP component shows that SHUM generates maximum spread over tropical regions, similar to SPPT (Fig. 3 in Zhu et al. 2019). Both SHUM and SPPT represent the model uncertainties related to convection. The sensitivity tests with and without SHUM show that the exclusion of SHUM can relieve the overdispersion issue of tropical low-level winds without a significant impact on other winds (not shown). SHUM is not utilized in the GEFSv12 ensemble forecasts.

The ensemble spread of zonal winds shows that the STTP produces a large ensemble spread over extratropical regions but a relatively small spread over the tropics (Fig. 2a). The new SP suite produces an ensemble spread similar to STTP, with a slight reduction in the winter hemisphere and an evident increase over the tropics. The contribution of SKEB resembles that of STTP, accounting for large uncertainties over baroclinic regions, while SPPT is responsible for the increased spread over the tropics (Figs. 2c,f).

c. *Ensemble size*

Increasing the ensemble size is desirable in ensemble prediction to improve the average skill and increase the reliability of the estimate of the forecast probability distribution (Buizza et al. 1999; Leutbecher 2018). However, it is always a cost–benefit issue as the ensemble size is proportional to the computational cost (Ma et al. 2012).

There were three ensemble members with two perturbed members and one control run when the first NCEP GEFS version became operational in December 1992. The ensemble size of GEFS, the model horizontal and vertical resolution, has increased gradually with the increase in computational resources over the past decades. The ensemble size has remained at 20 for the past 13 years since the GEFS version 8 was implemented in July 2007, until it increased to 30 in the latest version GEFSv12.

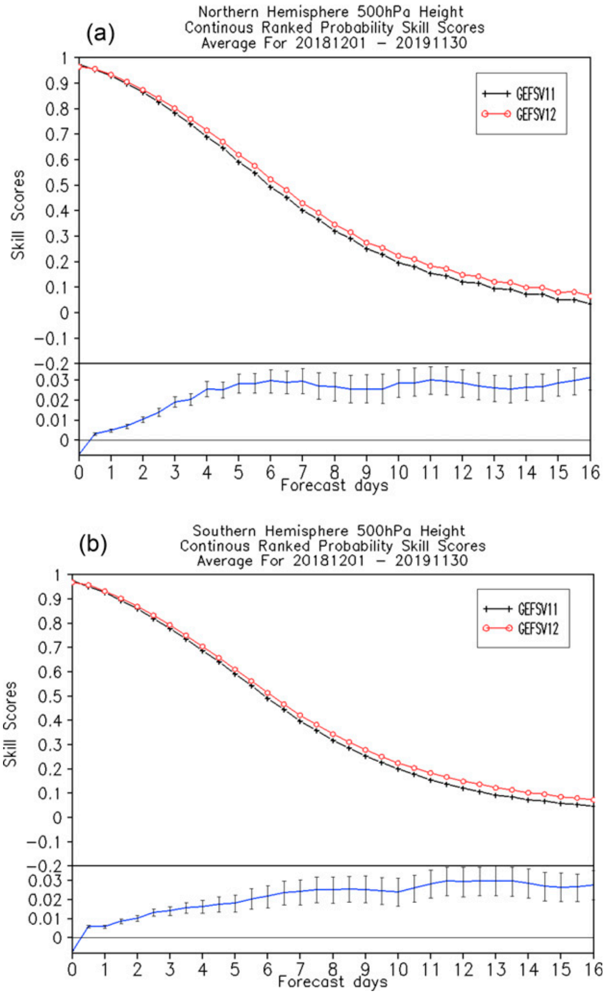

FIG. 9. The CRPSS for the 500-hPa geopotential height in the (a) NH and (b) SH. The black curves represent GEFSv11, and the red curves represent GEFSv12. The lower graphs show the CRPSS differences and bootstrap significance tests. The difference is significant at the 95% confidence level when the bars do not overlap with the horizontal axis through zero.

The benefits of increasing the number of ensemble members from 20 to 30 were generally observed in all evaluation metrics, but the degree of improvement varied depending on the metric. For 500-hPa geopotential height, the changes in the ensemble-mean forecasts to RMSE and PAC are very minor (not shown), while there is a slight improvement in CRPSS (Fig. 3a), as well as in BS and CRPS (not shown). Previous studies suggest that BS and CRPS of an *M*-member ensemble decrease asymptotically as $1 + M^{-1}$ relates to the score of an infinite-sized ensemble for perfect reliability

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

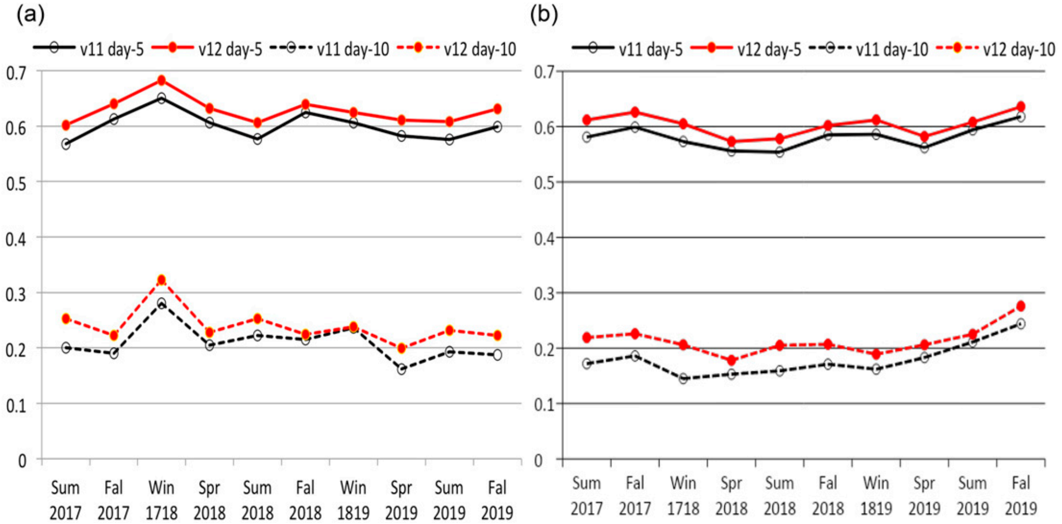

FIG. 10. The time series of the seasonal mean CRPSS for 500-hPa geopotential height at 5- and 10-day forecast lead times in the (a) NH and (b) SH. The black curves represent GEFSv11, and the red curves represent GEFSv12. The solid lines represent 5-day forecasts while the dashed ones are 10-day forecasts.

(Richardson 2001; Leutbecher 2018). According to this estimation, the CRPS/BS in a perfectly reliable 20-member (30-member) ensemble is 5% (3.3%) worse than the CRPS of an ensemble with an infinite number of members. The limited improvement in CRPS/BS is consistent with the estimates.

The improvement in the ROC score is evident with 30 ensemble members compared to 20 members (Fig. 3b). ROC computes the hit rate (probability of detection) and false alarm rate (probability of false detection) in each category. Higher ROC scores mean a higher hit rate and lower false alarm rate, indicating a better ability to distinguish between the occurrence and nonoccurrence of an event. ROC represents the "potential skill" of probability forecasts as it is independent of forecast bias. The increased ROC scores suggest that the impact of the ensemble size increases from 20 to 30 on the potential skill is considerable.

For precipitation probability forecasts over CONUS, there is an evident improvement in terms of BSS for precipitation probability forecasts at all thresholds including 1 and 20 mm (Fig. 4), as well as 5 and 10 mm (not shown). Leutbecher (2018) studied the dependence of the quantile score on ensemble size and suggested that for extreme probability levels close to zero or close to one, the convergence of quantile score with ensemble size is much slower than for the CRPS. Our results are consistent with Leutbecher's (2018) conclusion. The precipitation forecasts as low-probability events could benefit considerably from even a limited increase of ensemble size from 20 to 30 members.

There is only a slight improvement in the ensemble-mean forecast of TC tracks. The mean track error is reduced by around 3% in the 7-day forecast. Limited improvement in the ensemble-mean track forecast is likely due to the minimal increase of ensemble spread of the TC track (not shown). Adding more members does not improve the ensemble mean, as the forecasts from ensemble members are just clones of the same forecast.

d. GFDL cloud microphysics

GFDL MP is a single-moment cloud microphysics with five prognostics cloud species: liquid, ice, snow, graupel, and rain more based on the Lin–Lord–Krueger cloud microphysics (Lin et al. 1983; Lord et al. 1984; Krueger et al. 1995) with substantial development (Zhou et al. 2019). The Zhao–Carr MP only has one prognostic cloud species: total cloud water. The forecast performance with the GFDL MP and Zhao–Carr MP is compared in the GEFSv12 configuration. The PAC scores of 500-hPa geopotential height in both Northern and Southern Hemispheres (NH and SH) are better at all lead times with the GFDL MP than the Zhao–Carr MP (Fig. 5). The improvement in the SH is more evident than in the NH.

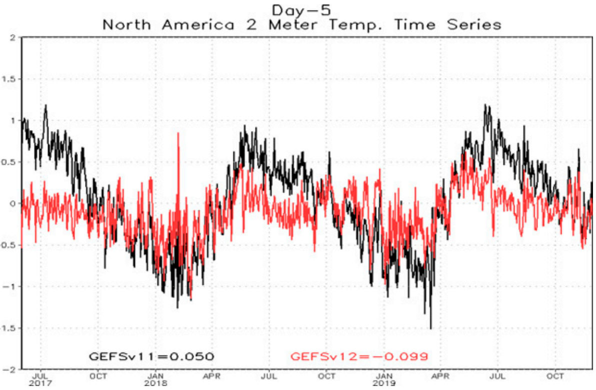

FIG. 11. The time series of temperature bias over CONUS. The black curves represent GEFSv11, and the red curves represent GEFSv12.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

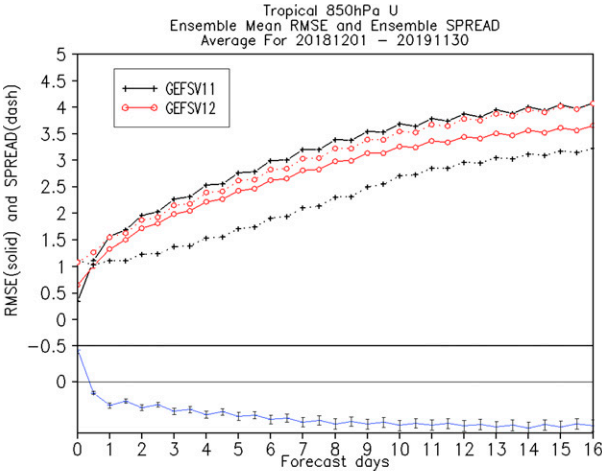

FIG. 12. As in Fig. 8, but for 850-hPa zonal wind over tropics. The black curves represent GEFSv11, and the red curves represent GEFSv12. The difference is significant at the 95% confidence level when the bars do not overlap with the horizontal axis through zero.

Interestingly, the new MP also significantly reduces the low-level warm bias in this testing period, from March 2017 to June 2017 (not shown). The TC track forecasts have mixed results as the ensemble-mean track forecasts are generally better over the Atlantic basin but slightly worse over the eastern Pacific, based on the TC forecasts for the 2016 and 2017 hurricane seasons (not shown).

e. *Horizontal advection options*

Weak biases of TC intensity are more evident in GEFSv12 than in GEFSv11 (Fig. 6), even though the former has a higher horizontal resolution. Following a suggestion from GFDL, the horizontal advection option (referred to as HORD) was changed. This modification results in a reduced TC intensity bias and ensemble-mean intensity forecast error up to 4 days. A very slight degradation is observed in the large-scale forecasts (not shown). It is considered acceptable considering its benefit on the TC intensity forecasts.

In the FV3, subgrid reconstruction from the cell-interface values of these variables is required to compute fluxes across cell interfaces for FV3 advection schemes. PPM is used for subgrid reconstruction following the formula (1.9) in Colella and Woodward (1984) with fourth-order accuracy:

$$a_{j+(1/2)} = \frac{7}{12}(a_j + a_{j+1}) - \frac{1}{12}(a_{j+2} + a_{j-1}), \qquad (1)$$

Parameter a_j is the cell-mean value of cell (j) and $a_{j+(1/2)}$ is defined as the interface value at the right-side interface value

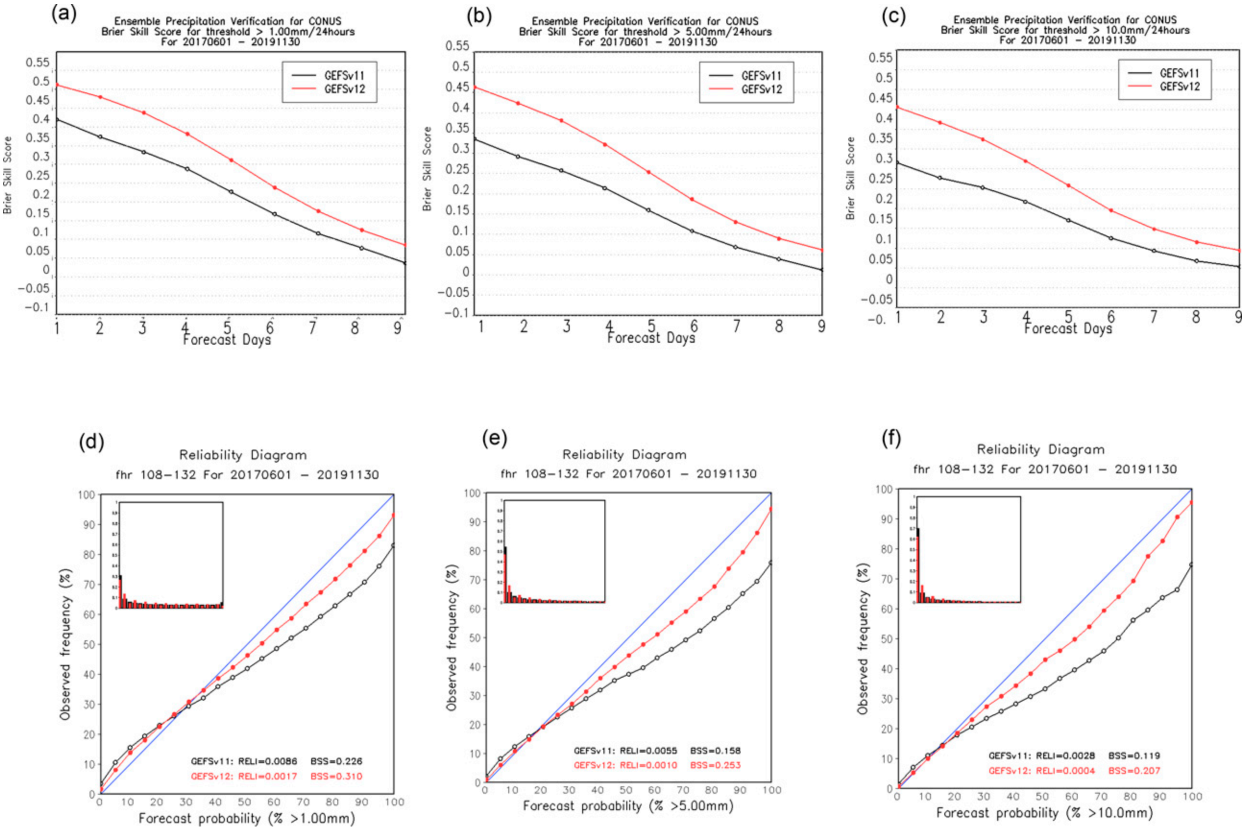

FIG. 13. BSS for the ensemble-mean precipitation greater than (a) 1, (b) 5, and (c) 10 mm day^{-1} for the period from June 2017 to November 2019. (d)–(f) As in (a)–(c), but for the reliability scores. The black curves represent GEFSv11, and the red curves represent GEFSv12. The bar chart in the upper left of the reliability diagrams shows the number of times that each probability value was predicted.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

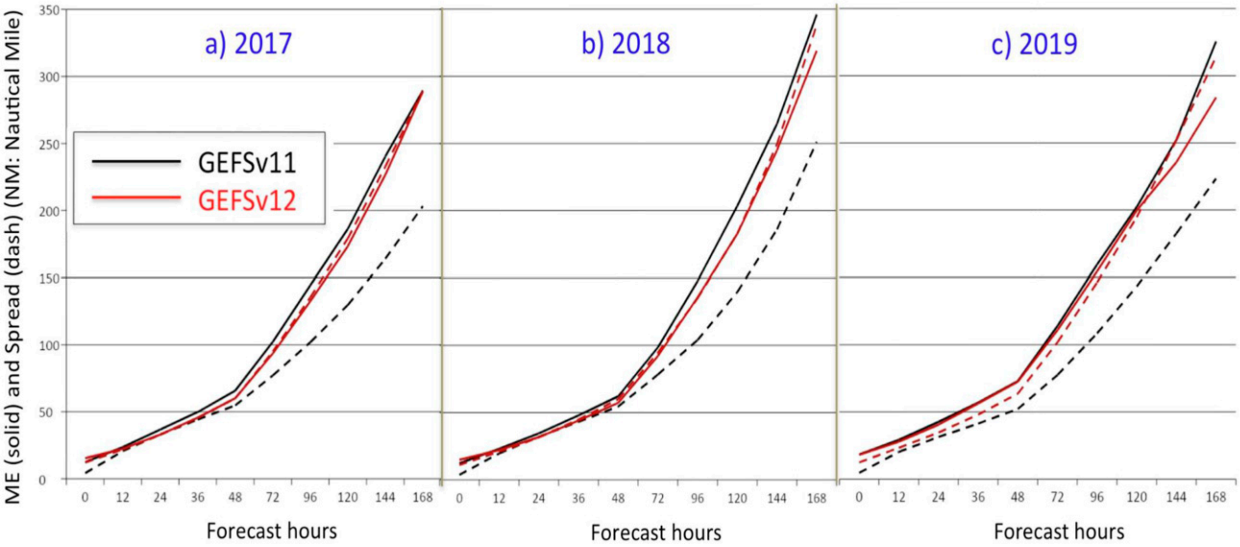

FIG. 14. The averaged ensemble-mean error of the tropical cyclone track forecast (solid) and ensemble spread (dotted) for all cases over the Atlantic, EP, and WNP basin in (a) 2017, (b) 2018, and (c) 2019. The black curves represent GEFSv11, and the red curves represent GEFSv12.

of cell (j), as well as the left-side interface value of cell ($j + 1$) by continuity. Several methods are available to further modify the interface value to prevent nonphysical oscillation. In an "unlimited" scheme with a weak $2\Delta x$ filter (referred to as HORD5), the following condition is evaluated to guarantee that interface values on opposite sides of a cell are either greater or less than the cell-mean value thus no new extrema is created:

$$\left(a_{j+(1/2)} - a_j\right)\left(a_{j-(1/2)} - a_j\right) < 0\,, \tag{2}$$

in which $a_{j-(1/2)}$ and $a_{j+(1/2)}$ are interface values at the left and right side of the cell (j) and a_j is as in (1). Otherwise, a first-order upwind flux is applied. Another method uses a stronger $2\Delta x$ filter (referred to as HORD6):

$$3\left|a_{j-(1/2)} + a_{j+(1/2)} - 2a_j\right| < \left|a_{j-(1/2)} - a_{j+(1/2)}\right|, \tag{3}$$

which not only filters $2\Delta x$ signals but also limits the slope steepness of reconstructed variables.

The monotonicity constraint is another well-known method to keep an originally monotonic distribution monotonic after advection. One monotonic option in FV3 is originally derived by Colella and Woodward (1984) and optimized by Lin (2004), with reduced numerical diffusion and improved computational efficiency (referred to as HORD8). This method is used in GEFSv12 for the tracer advection.

A finite-volume one-dimensional scalar differential scheme is carried out to better understand the impact of advection options on model performance. The transportation of a rectangular wave and Gaussian wave with periodic boundary conditions and mean flow from left to right is compared. There are no differences among these three approaches for the advection of a continuous scalar (not shown). The results from the advection of a rectangular wave show that HORD8 is most diffusive, while HORD5 is the least (Fig. 7). There are more evident overshoots and undershoots in HORD5 than in HORD6 and no overshoots in HORD8, due to strictly monotonic constraints. Less diffusivity in HORD5 could be responsible for more intense hurricanes than in HORD6. Overshoots at discontinuity regions such as the areas with convective rainbands and TC eyewalls could favor deep moist updrafts, thus leading to more intense TCs. Our results are consistent with Gao et al.'s (2021) study. Their investigation shows that hurricane intensity and structure are sensitive with the horizontal tracer advection scheme. By comparing two schemes, a monotonic scheme (HORD8) and HORD5 with additional positive-definite constraints, they found that replacing the tracer advection option HORD8 with HORD5 leads to stronger storms.

Given that HORD5 is a less diffusive scheme, the kinetic energy dissipation is larger than in HORD6. The parameter that controls the amplitude of SKEB perturbations is reduced around 1/3 when HORD6 is changed to HORD5 to avoid overspread since SKEB perturbations are based on the estimate of the kinetic energy dissipation.

5. Verification of GEFSv12 retrospective runs

After the configuration of GEFSv12 is determined based on extensive testing and verification, retrospective runs for more than 2 years are carried out to support the validation for all GEFS stakeholders. The performance of the retrospective runs is compared with that of the GEFSv11 for the forecasts up to 16 days and the GEFS SubX for the extended forecasts. With the limited computing resources, the retrospective runs provided 30-member ensemble forecasts up to 16 days for one year from December 2018 to November 2019, and up to 10-day forecasts for one year from June 2017 to November

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

2018. All retrospective runs are initialized at 0000 UTC daily, but the forecasts initialized at 1200 UTC are added during the hurricane seasons of NH (July–October) to increase the sample size of tropical cyclone forecasts. To evaluate MJO forecast skills, the retrospective runs initialized at 0000 UTC extend to 35 days for the cases from May 2017 to April 2018 every 7 days. A comprehensive evaluation is performed, and the major conclusions are summarized.

a. *Grid-to-grid verification*

The performance of GEFSv12 is significantly better overall than GEFSv11 concerning the verification metrics included in the GEFS verification package. The RMSE for 500-hPa geopotential height in both the NH and SH is reduced significantly, and there is a better spread–error relationship in GEFSv12 compared with GEFSv11 (Fig. 8). The underdispersion is evident in the NH in both GEFSv12 and GEFSv11, but the application of the new stochastic physics suites results in a large spread at all forecast lead times. The spread of GEFSv11 in the SH is underdispersive in the first week but overdispersive in the second week. GEFSv12 has an increased ensemble spread at short lead times but remains underdispersive at all lead times.

A metric that measures the improvement of the ensemble-mean forecasts is the forecast lead time of skillful forecasts with an anomaly correlation larger than 0.6 in terms of the 500-hPa geopotential height. By comparing the corresponding anomaly correlation between GEFSv12 and GEFSv11, we found that the skillful forecasts in GEFSv12 extend to 10 days from the 9.7 days in GEFSv11 in the NH.

GEFSv12 outperforms GEFSv11 in the probability forecasts with improved CRPSS, ROC, and BSS in all forecast lead times (only CRPSS shown in Fig. 9). Figure 10 shows the time series of 3-month mean CRPSS for the 500-hPa geopotential height in the NH and SH during the 2-yr retrospective periods. The improved performance is consistent over time in terms of the monthly averaged CRPSS.

The improvement in the forecasts of 850- and 250-hPa wind fields in the NH is also evident for CRPSS (not shown). The skillful forecasts with CRPSS greater than a threshold of 0.3 extended about 0.5–0.6 days for the forecasts of winds at 850 and 250 hPa.

Large warm/cold biases of 2-m temperature over North America (NA) in the warm/cold season have been noted in the GEFSv11 (Zhou et al. 2017). These biases with seasonal variations are substantially reduced in GEFSv12, especially in the warm seasons (Fig. 11). The reduced warm bias is likely associated with the replacement of the Zhao–Carr MP with the GFDL MP, since a similar improvement was observed in the corresponding sensitivity tests (not shown). The 2-m temperature bias in winter is reduced, but a slight cold bias remains in GEFSv12.

The forecasts of winds, temperature, and geopotential height are significantly improved in the tropics (not shown), as in the NH and SH. A notable change is the substantially increased ensemble spread ascribed to the application of SPPT. The large underdispersion in GEFSv11 is substantially reduced, except where it becomes overdispersive in the GEFSv12 in terms of low-level winds. Figure 12 shows that the RMSE of 850-hPa zonal wind is reduced significantly, but the ensemble spread is larger than RMSE in GEFSv12.

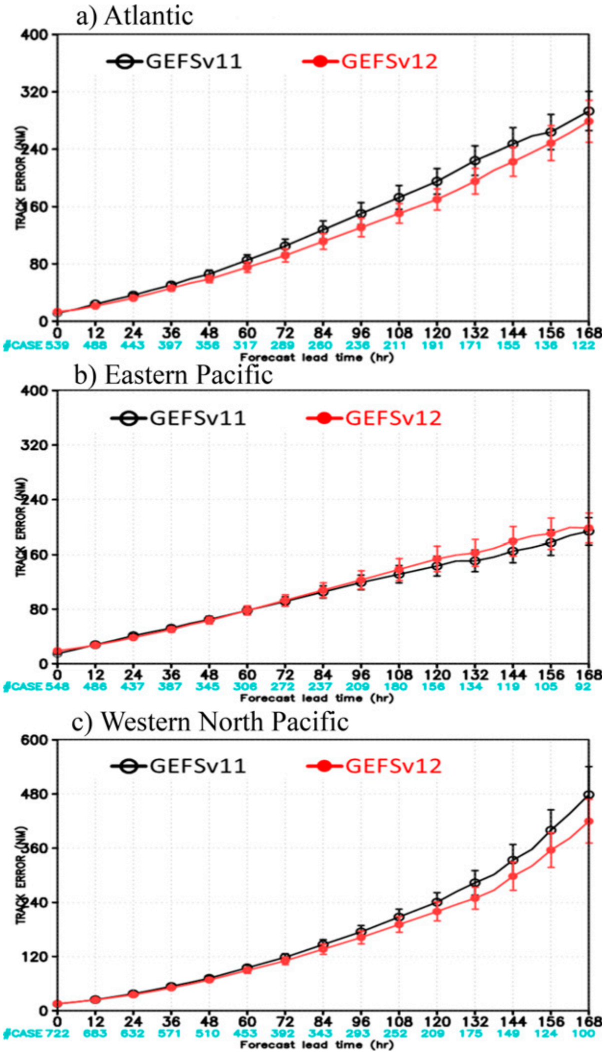

FIG. 15. The tropical cyclone track forecast error of the ensemble-mean forecasts for the tropical cyclones over the (a) Atlantic basin, (b) EP, and (c) WNP averaged from 2017 to 2019. The blue is for GEFSv11, and the red is for GEFSv12. The bars represent the difference is significant at the 95% confidence level.

b. *Precipitation*

The precipitation forecasts are only verified over CONUS. GEFSv12 generally outperforms GEFSv11 in terms of precipitation probability forecasts. The BS and reliability scores are always better in GEFSv12 than in GEFSv11. The BS measures mean squared error in probability space. Figure 13 shows that the BS of GEFSv12 is smaller at all lead times than those of GEFSv11. Reliability diagrams provide information about probability forecast bias. A reliability curve along the diagonal line represents no probability forecast bias. The comparison shows that the reliability curves of GEFSv12 are generally closer to the diagonal line than those from

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

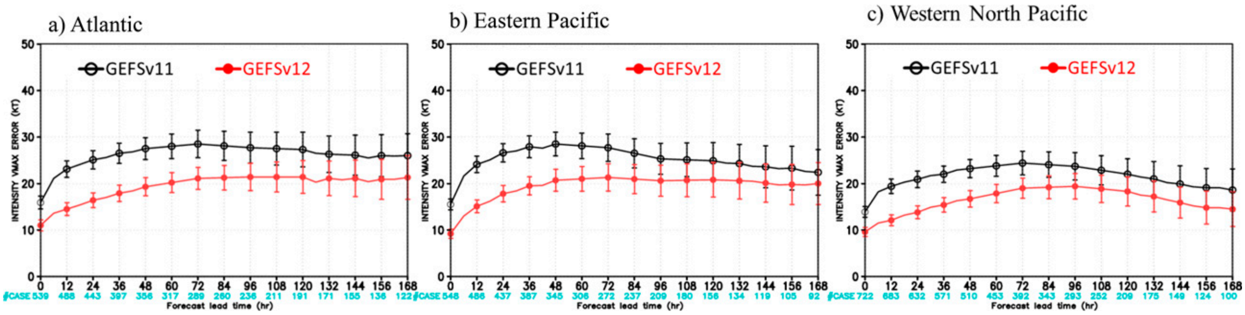

FIG. 16. As in Fig. 15, but for tropical cyclone intensity forecast error.

GEFSv11 for both high and low probability categories. The probabilities are overestimated/underestimated as the reliability curve is located at the right/left side of the diagonal line. GEFSv12 has much more reliable ensemble forecasts than GEFSv11.

c. *Tropical cyclone track and intensity forecasts*

The tropical cyclone forecasts over the Atlantic, eastern Pacific (EP), and western North Pacific (WNP) were verified by year and basin for the period of the retrospective runs. Figure 14 shows that the RMSE of tropical cyclone track forecasts is reduced each year. The ensemble spread of the forecasted tracks is considerably underdispersive in GEFSv11 while it is much closer to the RMSE in the GEFSv12. It means that GEFSv11 generally underestimates the uncertainties of track forecasts while GEFSv12 has a much better spread–error relationship. The improved spread–error relationship is primarily due to the upgrade of the model stochastic schemes. Note that the performance of tropical cyclone track forecasts in GEFSv12 varies in the three basins. The track forecasts are improved in the Atlantic and WNP but are slightly degraded over the EP for days 3–7 (Fig. 15).

The intensity forecasts are significantly improved over all basins (Fig. 16) as the tropical storms are more intense in the new system than in the old version. The selection of the less diffusive advection option (HORD5) contributes to more intense TCs in GEFSv12.

d. *MJO forecasts*

The forecast skill of MJO, which is considered to be the most important predictability source for the subseasonal time scale in the tropics, was verified using the real-time multivariate MJO (RMM1 and RMM2) of Wheeler and Hendon (2004). GEFSv12 is compared against the experimental SubX GEFS. The comparison shows that the skillful MJO prediction with the Wheeler–Hendon MJO index > 0.6 extends for 2 days from the SubX GEFS to GEFSv12 for the 1 year from 2017 to 2018 (Fig. 17a). The improvement in MJO forecast skill is primarily from the improved forecasts of the OLR components (Fig. 17b), while the performance of zonal winds at 200- and 850-hPa pressure levels (referred to as U200 and U850) are similar in these two systems (Figs. 17c,d). Note that we only have a very limited sample size for 35-day forecasts in the retrospective runs due to the limitation of computational

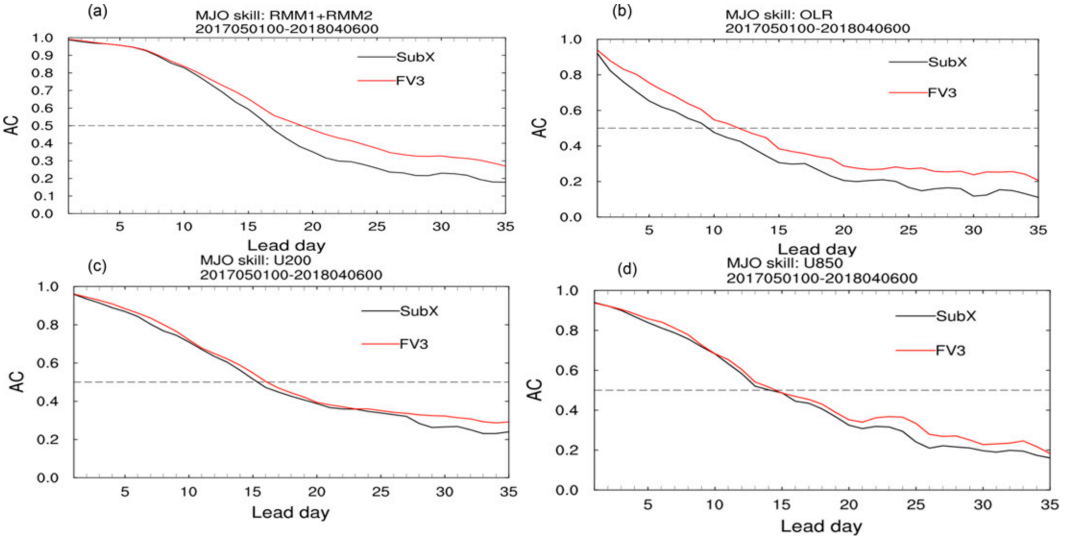

FIG. 17. (a) The MJO prediction skill scores for GEFSv12 (red line) and SubX (black line). The contributions of (b) OLR, (c) U200, and (d) U850 to the MJO prediction skill.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

resources. An objective verification of MJO forecast skills of GEFSv12 should be obtained from the corresponding reforecast dataset.

6. Summary and discussion

The GEFS is upgraded to version 12 following the overall modeling development strategy and implementation plan developed by NOAA. Major upgrades from the GEFSv11 include 1) the switch from a GSM model to an FV3-based model, 2) extending the forecast length from 16 days to 35 days to cover subseasonal time scale forecasts, 3) increasing the horizontal resolution from 34 to 25 km, 4) increasing the number of ensemble members from 20 to 30, 5) replacing the Zhao–Carr MP scheme with the GFDL microphysics scheme, and 6) using SKEB and SPPT as the model uncertainty schemes to replace STTP.

The comparison between GEFSv12 and v11 is performed based on more than 2 years of retrospective runs. The upgraded GEFS presents significant improvement in many aspects. GEFSv12 outperformed GEFSv11 in terms of the forecasts of traditional forecast variables, rainfall, tropical cyclone track, and intensity, as well as MJO. The improvement also presents itself in different verification metrics, including the ensemble mean forecast skill scores of RMSE and PAC and probability skill scores such as BSS, CRPSS, and ROC. QPF over the CONUS is much more reliable and more skillful. SPPT and SKEB improve the error–spread relationship compared with STTP, especially in the tropics. Even though the underdispersion of temperature and wind fields are generally common in both versions beyond 7 days in the NH, the spread of GEFSv12 increased slightly. The forecast errors of tropical cyclone tracks are reduced over the WNP and Atlantic but slightly increased over the EP. GEFS has a long history of complaints from users that the ensemble forecasts of tropical cyclone tracks are too underdispersive. GEFSv12 has evident improvement in the spread–error relationship in terms of tropical cyclone track forecasts. GEFSv12 also reduces the weak intensity bias of tropical cyclones and has smaller intensity errors. In addition, the new system reduces the 2-m temperature warm bias in warm seasons and cold bias in cold seasons over North America.

Our study has focused on the performance evaluation for the upgraded GEFS as an integrated system. It is impossible to demonstrate the contribution of each upgraded component to the overall improved forecast performance. However, we can obtain some clues from the sensitivity experiments we performed during the development period. For example, the GFDL-MP scheme likely plays an important role in reducing the 2-m temperature bias. SPPT efficiently increases the ensemble spread of GEFS, especially over the tropics, which is essential in increasing the spread of forecast tropical cyclone tracks. The use of a less diffusive advection scheme leads to generally more intense tropical cyclones, thus reducing tropical cyclone intensity forecast errors.

A long training dataset of reforecasts is highly desirable for a subseasonal forecast system to remove systematic errors and reshape the predicted probability distribution. A global reanalysis dataset from 2000 to 2019 and a reforecast dataset from 1990 to 2019 have been created and have been available to the public. Nevertheless, the subseasonal forecasts in this study are verified based on the 1-yr retrospective run without the removal of systematic bias. A more comprehensive verification of the MJO forecast skills with 20-yr reforecast data will be presented in another paper.

Large-scale anomalies in the initial state with slowly varying processes such as upper-ocean heat content and sea ice are the major sources of predictability at subseasonal time scales. However, GEFSv12 is still an atmosphere-only system. Two-tiered SSTs using calibrated CFS SST forecasts and a persistent method for sea ice are used as the lower boundary forcing for the atmosphere, which limits the predictability in subseasonal time scales. A fully coupled model with FV3-based atmospheric, ocean (GFDL Modular Ocean Model MOM6), sea ice (CICE), and land (Noah land surface model) components are being developed for implementation in the next upgrade cycle (GEFSv13).

Acknowledgments. The authors thank the FV3 GFS group, especially Fanglin Yang, Jongil Han, Ruiyu Sun, Weizhong Zheng, Shrinivas Moorthi, Daryl Kleist, Rahul Mahajan, Jack Kain, Jeffrey Whitaker, and Tom Hamill for setting the model configuration. The authors also appreciate the help from the GFDL FV3 group, especially Lucas Harris, Linjiong Zhou, and Xi Chen. We also thank Catherine Thomas and Xingren Wu for their careful EMC internal review and Mary Hart for aiding the editorial review of the manuscript.

REFERENCES

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.*, **66**, 603–626, https://doi.org/10.1175/2008JAS2677.1.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Colella, P., and P. R. Woodward, 1984: The Piecewise Parabolic Method (PPM) for gas-dynamical simulations. *J. Comput. Phys.*, **54**, 174–201, https://doi.org/10.1016/0021-9991(84)90143-8.

Gao, K., L. Harris, L. Zhou, M. A. Bender, and M. J. Morin, 2021: On the sensitivity of hurricane intensity and structure to horizontal tracer advection schemes in FV3. *J. Atmos. Sci.*, **78**, 3007–3021, https://doi.org/10.1175/JAS-D-20-0331.1.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., 1999: Hypothesis tests for evaluation numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull.*

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

Amer. Meteor. Soc., **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, and Coauthors, 2021: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, https://doi.org/10.1175/WAF-D-10-05038.1.

——, W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Wea. Forecasting*, **32**, 2005–2017, https://doi.org/10.1175/WAF-D-17-0046.1.

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

——, X. Chen, L. Zhou, and J.-H. Chen, 2020a: The nonhydrostatic solver of the GFDL finite-volume cubed-sphere dynamical core. Tech. Memo. 2020-003, Geophysical Fluid Dynamics Laboratory, 6 pp., https://repository.library.noaa.gov/view/noaa/27489.

——, and Coauthors, 2020b: GFDL SHiELD: A unified system for weather-to-seasonal prediction. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002223, https://doi.org/10.1029/2020MS002223.

Hou, D., Z. Toth, and Y. Zhu, 2006: A stochastic parameterization scheme within NCEP Global Ensemble Forecast System. *18th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., 4.5, https://ams.confex.com/ams/Annual2006/techprogram/paper_101401.htm.

——, ——, ——, and W. Yang, 2008: Impact of a stochastic perturbation scheme on NCEP Global Ensemble Forecast System. *19th Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/88Annual/techprogram/paper_134165.htm.

——, and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Ji, M., and F. Toepfer, 2016: Dynamical core evaluation test report for NOAA's Next Generation Global Prediction System (NGGPS). NOAA IR ID 18653, 93 pp., https://doi.org/10.25923/ztzy-qn82.

Krueger, S. K., Q. A. Fu, K. N. Liou, and H. N. S. Chin, 1995: Improvements of an ice-phase microphysics parameterization for use in numerical simulations of tropical convection. *J. Appl. Meteor.*, **34**, 281–287, https://doi.org/10.1175/1520-0450-34.1.281.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.*, **121**, 2030–2045, https://doi.org/10.1175/1520-0493(1993)121<2030:AISOHM>2.0.CO;2.

——, ——, R. E. Tuleya, and R. J. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.*, **123**, 2791–2801, https://doi.org/10.1175/1520-0493(1995)123<2791:IITGHP>2.0.CO;2.

Leutbecher, M., 2018: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145** (Suppl.), 107–128, https://doi.org/10.1002/qj.3387.

Li, W., and Coauthors, 2018: Evaluating the MJO prediction skill from different configurations of NCEP GEFS extended forecast. *Climate Dyn.*, **52**, 4923–4936, https://doi.org/10.1007/s00382-018-4423-9.

Lin, S.-J., 1997: A finite-volume integration method for computing pressure gradient force in general vertical coordinates. *Quart. J. Roy. Meteor. Soc.*, **123**, 1749–1762, https://doi.org/10.1002/qj.49712354214.

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

——, 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-Langrangian shallow-water model on the sphere. *Quart. J. Roy. Meteor. Soc.*, **123**, 2477–2498, https://doi.org/10.1002/qj.49712354416.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snowfield in a cloud model. *J. Climate Appl. Meteor.*, **22**, 1065–1092, https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2.

Liu, Q., S. J. Lord, N. Surgi, Y. Zhu, R. Wobus, Z. Toth, and T. Marchok, 2006: Hurricane relocation in global ensemble forecast system. *27th Conf. on Hurricanes and Tropical Meteorology*, Monterey, CA, Amer. Meteor. Soc., P5.13, https://ams.confex.com/ams/pdfpapers/108503.pdf.

Lord, S. J., H. E. Willoughby, and J. M. Piotrowicz, 1984: Role of a parameterized ice-phase microphysics in an axisymmetric, nonhydrostatic tropical cyclone model. *J. Atmos. Sci.*, **41**, 2836–2848, https://doi.org/10.1175/1520-0469(1984)041<2836:ROAPIP>2.0.CO;2.

Lott, F., and M. J. Miller, 1997: A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quart. J. Roy. Meteor. Soc.*, **123**, 101–127, https://doi.org/10.1002/qj.49712353704.

Ma, J., Y. Zhu, D. 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.

McCormack, J. P., S. D. Eckermann, D. E. Siskind, and T. J. McGee, 2006: CHEM2D-OPP: A new linearized gas-phase ozone photochemistry parameterization for high-altitude NWP and climate models. *Atmos. Chem. Phys.*, **6**, 4943–4972, https://doi.org/10.5194/acp-6-4943-2006.

——, K. W. Hoppel, and D. E. Siskind, 2008: Parameterization of middle atmospheric water vapor photochemistry for high-altitude NWP and data assimilation. *Atmos. Chem. Phys.*, **8**, 7519–7532, https://doi.org/10.5194/acp-8-7519-2008.

Palmer, T. N., 1997: On parametrizing scales that are only somewhat smaller than the smallest resolved scales, with application to convection and orography. *Workshop on New Insights and Approaches to Convective Parameterization*, Reading, United Kingdom, ECMWF, 328–337, https://www.ecmwf.int/node/11493.

——, 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Quart. J. Roy.*

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC

Meteor. Soc., **127**, 279–304, https://doi.org/10.1002/qj.49712757202.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multi-model subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

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

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, https://doi.org/10.1002/qj.49712757715.

Saha, S., and Coauthors, 2014: The NCEP climate forecast system version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, https://doi.org/10.1256/qj.04.106.

Tolman, H. L., 2016: User manual and system documentation of WAVEWATCH III version 5.16. NOAA/NWS/NCEP/MMAB Tech. Note 329, 361 pp., https://polar.ncep.noaa.gov/waves/wavewatch/manual.v5.16.pdf.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

——, ——, and Y. Zhu, 2006: The attributes of forecast systems. *Predictability of Weather and Climate*, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 584–595.

Wang, J., and Coauthors, 2018: The implementation of NEMS GFS Aerosol Component (NGAC) Version 2.0 for global multispecies forecasting at NOAA/NCEP – Part 1: Model descriptions. *Geosci. Model Dev.*, **11**, 2315–2332, https://doi.org/10.5194/gmd-11-2315-2018.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>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. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: Comparison of the ensemble transform and the ensemble Kalman filter in the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074, https://doi.org/10.1175/WAF-D-16-0109.1.

——, ——, ——, Y. Luo, J. Peng, and D. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, Y., and Z. Toth, 2008: Ensemble-based probabilistic forecast verification. *19th Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., 2.2, https://ams.confex.com/ams/88Annual/webprogram/Paper131645.html.

——, X. Zhou, M. Peña, W. Li, C. Melhauser, and D. Hou, 2017: Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP Global Ensemble Forecasting System. *Wea. Forecasting*, **32**, 2159–2174, https://doi.org/10.1175/WAF-D-17-0093.1.

——, and Coauthors, 2018: Toward the improvement of subseasonal prediction in the National Centers for the Environmental Prediction Global Ensemble Forecast System. *J. Geophys. Res. Atmos.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.

——, W. Li, X. Zhou, and D. Hou, 2019: Stochastic representation of NCEP GEFS to improve sub-seasonal forecast. *Current Trends in the Representation of Physical Processes in Weather and Climate Models*, D. A. Randall et al., Eds., Springer, 317–328.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 02/23/23 07:53 PM UTC