

## Evaluation of Tropical Cyclone Forecasts in the Next Generation Global Prediction System

JAN-HUEY CHEN,<sup>a,b</sup> SHIAN-JIANN LIN,<sup>b</sup> LINJIONG ZHOU,<sup>b,c</sup> XI CHEN,<sup>b,c</sup> SHANNON REES,<sup>a,b</sup>  
MORRIS BENDER,<sup>b,c</sup> AND MATTHEW MORIN<sup>a,b</sup>

<sup>a</sup> *University Corporation for Atmospheric Research, Boulder, Colorado*

<sup>b</sup> *NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*

<sup>c</sup> *Program in Atmosphere and Ocean Sciences, Princeton University, Princeton, New Jersey*

(Manuscript received 27 June 2018, in final form 25 June 2019)


### ABSTRACT

A new global model using the GFDL nonhydrostatic Finite-Volume Cubed-Sphere Dynamical Core (FV3) coupled to physical parameterizations from the National Centers for Environmental Prediction's Global Forecast System (NCEP/GFS) was built at GFDL, named fvGFS. The modern dynamical core, FV3, has been selected for the National Oceanic and Atmospheric Administration's Next Generation Global Prediction System (NGGPS) due to its accuracy, adaptability, and computational efficiency, which brings a great opportunity for the unification of weather and climate prediction systems. The performance of tropical cyclone (TC) forecasts in the 13-km fvGFS is evaluated globally based on 363 daily cases of 10-day forecasts in 2015. Track and intensity errors of TCs in fvGFS are compared to those in the operational GFS. The fvGFS outperforms the GFS in TC intensity prediction for all basins. For TC track prediction, the fvGFS forecasts are substantially better over the northern Atlantic basin and the northern Pacific Ocean than the GFS forecasts. An updated version of the fvGFS with the GFDL 6-category cloud microphysics scheme is also investigated based on the same 363 cases. With this upgraded microphysics scheme, fvGFS shows much improvement in TC intensity prediction over the operational GFS. Besides track and intensity forecasts, the performance of TC genesis forecast is also compared between the fvGFS and operational GFS. In addition to evaluating the hit/false alarm ratios, a novel method is developed to investigate the lengths of TC genesis lead times in the forecasts. Both versions of fvGFS show higher hit ratios, lower false alarm ratios, and longer genesis lead times than those of the GFS model in most of the TC basins.

### 1. Introduction

Due to the hazardous impacts that tropical cyclones (TCs) have on human lives and economic activities, TC prediction has always been an important task for weather forecast agencies in many countries. In the United States, the Global Forecast System (GFS) operated by the National Centers for Environmental Prediction (NCEP) provides the front-line guidance for most severe weather events, including TCs. According to annual verification reports from

operational centers,<sup>1</sup> the GFS and other global models [e.g., the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) and the Met Office Unified Model (UM) from the United Kingdom] are among the top performers for TC track predictions during the most recent years (2015–17). However, the forecast of TC intensity has remained a challenge for all global models due to its requirements of high horizontal and vertical resolutions, and more advanced physical parameterizations (Bender et al. 2017; Hazelton et al. 2018).

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Jan-Huey Chen, jan-huey.chen@noaa.gov

<sup>1</sup> These include the National Hurricane Center (NHC; 2017, <http://www.nhc.noaa.gov/verification/verify3.shtml>) for the North Atlantic and northeast Pacific basins and the Joint Typhoon Warning Center (JTWC; 2017 <http://www.metoc.navy.mil/jtwc/jtwc.html?cyclone>) for the northwest Pacific basin, the north Indian Ocean, and the Southern Ocean.

A global general circulation model (GCM) with cloud resolving capability has been pursued for almost a decade (Satoh et al. 2008). For a weather or climate model, the resolution is largely restricted by its computational efficiency, which is highly dependent on the dynamical core. The Finite-Volume Cubed-Sphere dynamical core (FV3) has been developed in the National Oceanic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL) over the past 10 years. The hydrostatic “vertically Lagrangian” finite-volume dynamical core (Lin 2004) was used in the GFDL CM2.1 model (Delworth et al. 2006) on a  $2^\circ$  by  $2.5^\circ$  latitude–longitude grid. To enhance the model’s parallel computing efficiency and simulations in polar regions, the dynamical core was reformulated on the cubed-sphere grid (Putman and Lin 2007) for GFDL CM2.5 and AM3/CM3 models (Delworth and Zeng 2012; Donner et al. 2011). The hydrostatic FV3 has also been adopted in many climate models outside of GFDL [e.g., the National Aeronautics and Space Administration (NASA) Goddard Earth Observing System Model, version 5 (GEOS-5);<sup>2</sup> NASA GEOS-CHEM;<sup>3</sup> NASA Goddard Institute for Space Studies (GISS) Atmospheric General Circulation Model (ModelE);<sup>4</sup> and the National Center for Atmospheric Research (NCAR) Community Earth System Model (CESM);<sup>5</sup> currently under implementation].

The nonhydrostatic FV3 has not been used in a global GCM until recently. The first full-physics model to apply the nonhydrostatic FV3 is the GFDL 25-km High Resolution Atmospheric Model (HiRAM), which shows skillful TC seasonal predictions and faithfully captures the observed influence of intra seasonal oscillation (ISO) on TC activity over the Atlantic basin (Chen and Lin 2016; Gao et al. 2017). From late 2014 to early 2017, the Next Generation Global Prediction System (NGGPS) Dynamical Core Evaluation was hosted by the Office of Science and Technology Integration (STI) in NOAA. During this comprehensive evaluation, the nonhydrostatic FV3 demonstrated its accuracy, scalability, and computational efficiency.<sup>6</sup> This modern dynamical core is selected to be the “engine” for the next generation GFS model for NOAA’s National Weather Service (NWS). Benefiting from the

high adaptability of FV3, the future GFS model is expected to provide a great opportunity for the unification of weather and climate prediction systems.

Many FV3-powered GCMs have been used for TC simulations at GFDL, mainly focused on seasonal predictions [e.g., HiRAM (Zhao et al. 2010; Chen and Lin 2011, 2013, 2016) and the High-Resolution Forecast-Oriented Low Ocean Resolution model (HiFLOR; Murakami et al. 2016)]. Skillful forecasts of seasonal storm counts were achieved at model resolutions from 25 to 50 km in both atmospheric-only and ocean-coupled models. Pursuing TC forecasts on a short-term weather time scale, a global model should contain not only adequate resolutions, but also proper initializations and physical parameterizations for the intended resolution. During NGGPS Phase II, a new global model using the FV3 nonhydrostatic dynamical core coupled to the physics package from NCEP/GFS was built at GFDL, called fvGFS. This model demonstrated equivalent forecast skills to the operational GFS at 13 km resolution without adjusting any parameters in the GFS physics package.<sup>7</sup>

The development of fvGFS has continued at GFDL after the end of NGGPS Phase II. One of the major updates is to use the GFDL cloud microphysics scheme to replace the Zhao–Carr gridscale condensation and precipitation scheme (Moorthi et al. 2001) in the original GFS physics package. The GFDL cloud microphysics scheme has been used in HiRAM for seasonal predictions at 25-km resolution (Chen and Lin 2011, 2013, 2016) and climate simulations on the stretched grid configuration (Harris et al. 2016). With this upgraded cloud microphysics, fvGFS shows a significantly better forecast skill than the operational GFS (Zhou et al. 2019).

In this study, we investigate the performance of global TC forecasts in fvGFS based on 363 cases of 10-day forecasts in 2015. TC track and intensity forecast errors as well as genesis forecasts in the two fvGFS configurations are compared to those in the operational GFS. One configuration uses the identical model code and settings submitted to the NGGPS Phase II dynamical evaluation, with operational GFS physical parameterizations used in 2015. The other configuration replaces the Zhao–Carr scheme with the GFDL cloud microphysics scheme but keeps other schemes in the operational GFS physical package. The detailed configurations of the fvGFS, TC best track dataset, and basic numerical weather prediction (NWP) skills are described in section 2. Section 3 shows the results of TC track and intensity forecast errors in the fvGFS as well as in the operational

<sup>2</sup> <https://gmao.gsfc.nasa.gov/systems/geos5/>.

<sup>3</sup> <http://acmg.seas.harvard.edu/geos/>.

<sup>4</sup> <https://www.giss.nasa.gov/tools/modelE/>.

<sup>5</sup> <http://www.cesm.ucar.edu/>; the current version of CESM uses the older latitude–longitude finite-volume core, the predecessor to FV3.

<sup>6</sup> The verification report for the NGGPS Phase II dynamical core evaluation can be found on the website: [https://www.weather.gov/sti/stimodeling\\_nggps\\_implementation\\_atmdynamics](https://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics).

<sup>7</sup> Detailed analysis results from NCEP/Environmental Modeling Center (EMC) can be found on the website: <http://www.emc.ncep.noaa.gov/gmb/wx24fy/nggps/web/>.

GFS. TC genesis forecasts in both fvGFS and GFS are compared and investigated in section 4. A summary and discussion are in section 5.

## 2. Model, data, methodology, and basic NWP skill

The fvGFS was built during the NGGPS Phase II Dynamical Core Evaluation for testing the robustness of the dynamical core under a wide range of realistic atmospheric initial conditions. The GFS physics package provided by NCEP/EMC was implemented under the GFDL Flexible Modeling System (FMS) framework (Balaji 2012) with the GFDL Finite-Volume Cubed Sphere Dynamical Core (FV3). Beyond the hydrostatic version of FV3 (Lin 2004), a nonhydrostatic solver is also included, with a semi-implicit approach for treating the vertically propagating sound waves, which replaces the hydrostatic balance equation used in the hydrostatic solver. To achieve the nonhydrostatic approach, pressure and geopotential are replaced with the nonhydrostatic full pressure and the true geopotential in the finite-volume pressure gradient scheme (Lin 1997). The GFS physics package includes: simplified Arakawa–Schubert (SAS) convection (Pan and Wu 1995), Zhao–Carr gridscale condensation and precipitation (Moorthi et al. 2001), orographic and convective gravity wave drag (Kim and Arakawa 1995; Chun and Baik 1994), boundary layer vertical diffusion (Hong and Pan 1996), and RRTM radiation scheme (Clough et al. 2005).

A key feature of FV3 is that the resolution is highly flexible. To compare to the operational GFS with horizontal resolution about 13 km along the equator, the model configuration used in NGGPS phase II was C768 which has 768 by 768 finite-volume cells on each face of the cube and 63 vertical levels with a model top at 0.64 hPa. The performance of the 10-day TC forecasts was investigated based on 363 cases from 16 January 2015 to 16 January 2016, using the 0000 UTC initial time for each forecast. Three days in 2015, 20 January, 28 July, and 14 November, were excluded due to missing GFS operational forecast data for comparison. The initial conditions for all 363 10-day forecasts are remapped from the GFS analysis data (Chen et al. 2018, 2019). The sea surface temperatures (SSTs) are also provided by each date’s GFS analysis but remain fixed from the initial time throughout the duration of each 10-day forecast.

Based on the regulation of Dynamical Core Evaluation, the configuration of fvGFS for NGGPS phase II submission, called “FV3\_zc” in this study, did not include parameter tunings on any of the GFS physical schemes. Therefore, the only major difference between FV3\_zc and the operational GFS forecast, called “GFS” in this study, is the dynamical core. The second configuration of fvGFS uses the GFDL cloud microphysics scheme (Zhou et al. 2019) to replace the Zhao–Carr gridscale

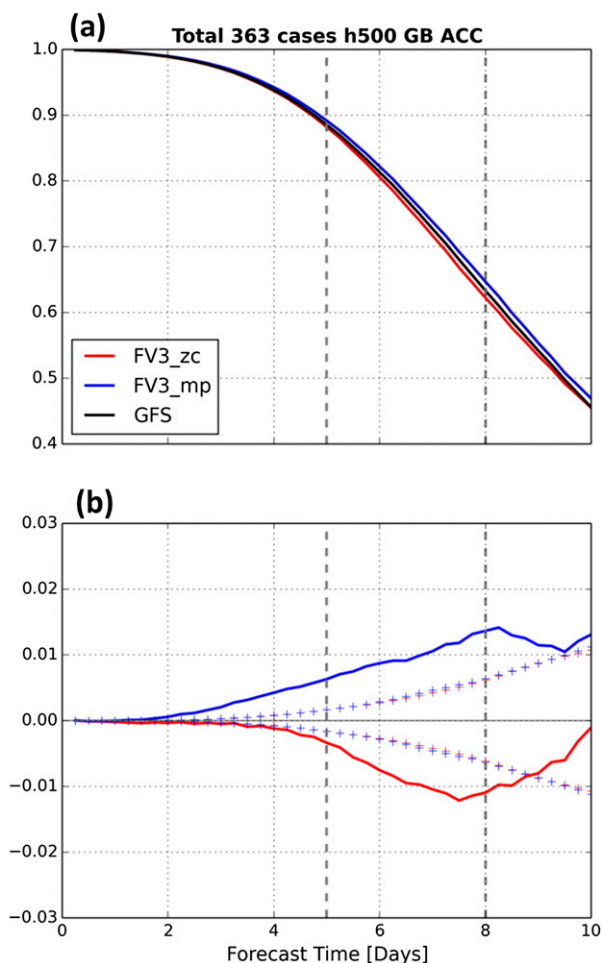


FIG. 1. (a) Global mean 500-hPa height ACC of 363 cases of 10-day forecasts for the operational GFS (black), FV3\_zc (red), and FV3\_mp (blue) based on the NCEP analysis data. (b) Differences of 500-hPa height ACC between FV3\_zc and GFS (red) and between FV3\_mp and GFS (blue). Vertical dashed lines (“- - -”) indicate day 5 and day 8. Dashed lines with “+ + +” represent the 95% confidence intervals. Positive values indicate an improvement relative to the GFS.

condensation and precipitation scheme (Moorthi et al. 2001), and includes some reconfiguration of the sponge layer in the dynamical core to better suppress shear instabilities in the upper stratosphere and mesosphere, but keeps other schemes in the GFS physics package unmodified. It is called “FV3\_mp” in this study. Figure 1a shows the global anomaly correlation coefficients (ACC) of 500-hPa height for the three sets of forecasts that are computed based on the NCEP analysis data. To highlight details of the ACC analysis, the ACC differences of the two fvGFS configurations compared to the operational GFS forecasts (GFS) are also shown in Fig. 1b. The fvGFS NGGPS Phase II submission version (FV3\_zc) only shows a slightly degraded forecast skill of 500-hPa ACC compared to GFS. It is a very encouraging result

TABLE 1. TCs observed during 16 Jan 2015 to 16 Jan 2016 based on the ATCF best track dataset. [Tropical storm (TS) and tropical depression (TD).]

North Atlantic basin		North Indian Ocean		North-central Pacific basin	
1	TS Ana	1	TS Ashobaa	1	TS Halola
2	TS Bill	2	TS Two	2	TS Iune
3	TS Claudette	3	TS Three	3	Hurricane Kilo
4	Hurricane Danny	4	Cyclone Chapala	4	Hurricane Loke
5	TS Erika	5	Cyclone Megh	5	TS Malia
6	Hurricane Fred			6	TS Niala
7	TS Grace			7	Hurricane Oho
8	TS Henri			8	TD Eight
9	TD Nine			9	TD Nine
10	TS Ida				
11	Hurricane Joaquin				
12	Hurricane Kate				
Northeast Pacific		Northwest Pacific		Southern Ocean	
1	Hurricane Andres	1	Typhoon Mekkhala	1	Cyclone Bansi
2	Hurricane Blanca	2	Typhoon Higos	2	TS Chedza
3	Hurricane Carlos	3	TS Bavi	3	TS Niko
4	TS Ela	4	Typhoon Maysak	4	TS Diamondra
5	Hurricane Dolores	5	TS Haishen	5	Cyclone Eunice
6	TS Enrique	6	Typhoon Noul	6	Cyclone Ola
7	TS Felicia	7	Typhoon Dolphin	7	TS Fundi
8	TD Eight	8	TS Kujira	8	Cyclone Lam
9	Hurricane Guillermo	9	Typhoon Chan-Hom	9	Cyclone Marcia
10	Hurricane Hilda	10	Typhoon-1 Linfa	10	TS Glenda
11	TD Eleven	11	Typhoon Nangka	11	TS Fifteen
12	Hurricane Ignacio	12	TS Twelve	12	TS Haliba
13	Hurricane Jimena	13	Typhoon Soudelor	13	Cyclone Pam
14	TS Kevin	14	TD Fourteen	14	Cyclone Nathan
15	Hurricane Linda	15	TS Molave	15	Cyclone Olwyn
16	TD Sixteen	16	Typhoon Goni	16	TS Reuben
17	Hurricane Marty	17	Typhoon Atsani	17	Cyclone Ikola
18	TS Nora	18	TS Etau	18	Cyclone Joalane
19	Hurricane Olaf	19	TS Vamco	19	TS Solo
20	Hurricane Patricia	20	Typhoon Krovanh	20	Cyclone Quang
21	TS Rick	21	Typhoon Dujan	21	TS Raquel
22	Hurricane Sandra	22	Typhoon Mujigae	22	TS One
		23	Typhoon Choi-Wan	23	TS Two
		24	Typhoon Koppu	24	TS Annabelle
		25	Typhoon CHampi	25	TS Tuni
		26	TD Twenty-Six	26	TS Bohale
		27	Typhoon In_Fa	27	Cyclone Ula
		28	Typhoon Melor	28	Cyclone Victor
		29	TD Twenty-Nine		

that after swapping the dynamical core, without applying any adjustments to the physical parameterizations, FV3\_zc achieved a comparable skill to GFS. Moreover, after replacing the cloud microphysics scheme, the forecast skill of FV3\_mp exceeds that of GFS after day 1 and remains slightly superior to the GFS throughout the duration of the 10-day forecast period (Fig. 1b).

### 3. TC track and intensity forecasts in fvGFS

There were 105 TCs observed during the period of 16 January 2015 to 16 January 2016. All TC names are

listed in Table 1 for the six active TC basins based on the classifications of the Automated Tropical Cyclone Forecast (ATCF) dataset;<sup>8</sup> for example, the North Atlantic Ocean, three basins of the northern Pacific Ocean (the northeast Pacific, the north-central Pacific, and the

<sup>8</sup>The ATCF best track data for the North Atlantic Ocean, and the northeast and north-central Pacific basins are downloaded from <https://ftp.nhc.noaa.gov/atcf/archive/>; the data for the northwest Pacific basin, the north Indian Ocean, and the Southern Ocean are from <https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks>.



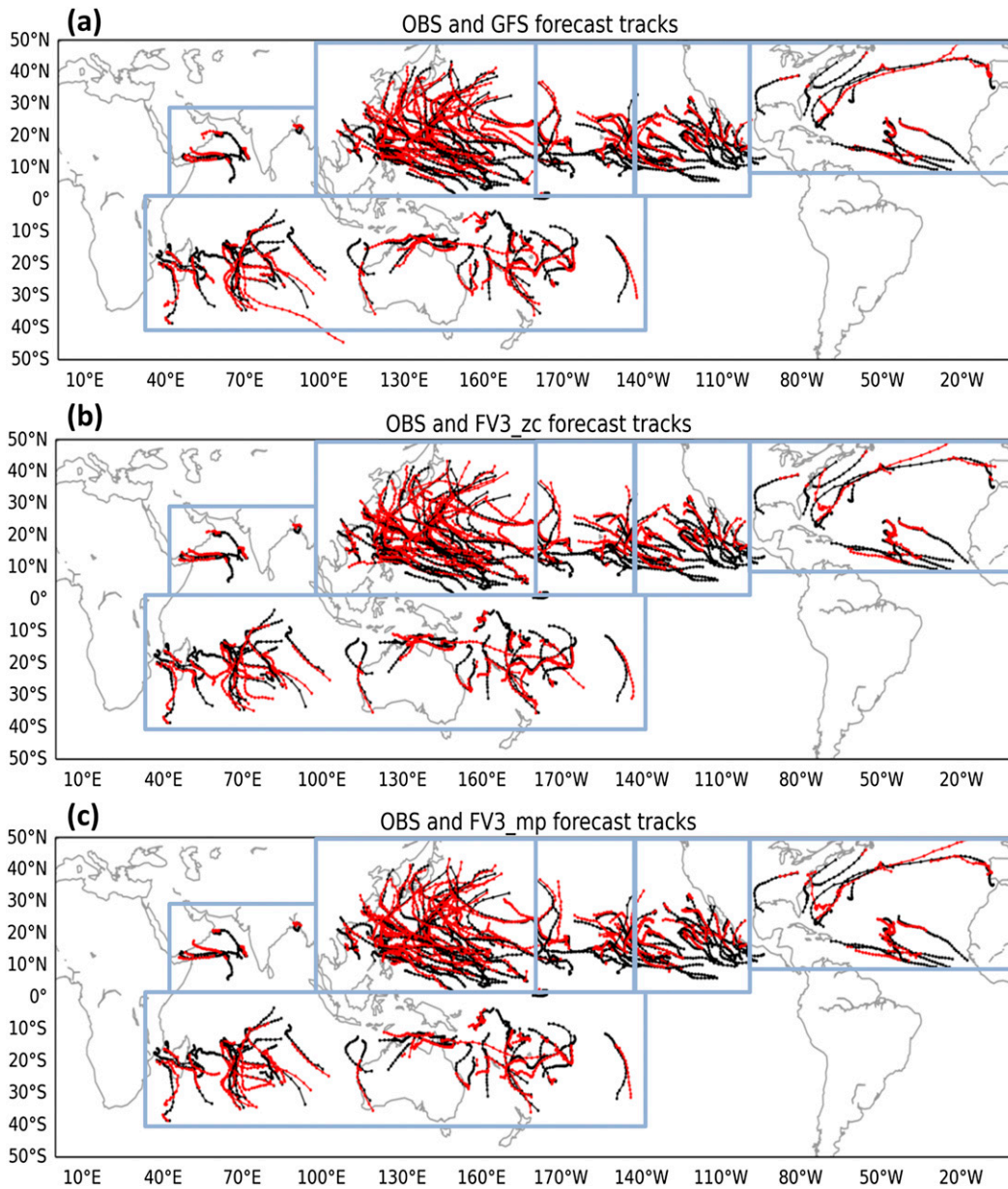


FIG. 2. Observed TC tracks from the ATCF best track datasets (black) and (a) NCEP/GFS operational (GFS), (b) FV3\_zc, and (c) FV3\_mp model forecasts. Forecasts (red) are plotted for only every 5 days for each case. The borderlines of the 6 TC regions based on the ATCF classification are indicated by light blue lines.

northwest Pacific), the north Indian Ocean, and the Southern Ocean (see light-blue borderlines in Fig. 2). To separate from storm genesis issues, results of TC forecast performance discussed in this section only consider the model forecasts that were initialized at or after the observed TC genesis time. We show results out to 168-h lead time, which is 2 days beyond the operational 5-day hurricane forecasts currently made by most operational centers.

The GFDL simple tracker (Harris et al. 2016) was adopted for tracking TCs in the FV3\_zc, the FV3\_mp, and the GFS model forecasts. For all of the results presented in this study, the three sets of forecasts were evaluated at their native resolution grid for a fair comparison. The errors of TC track and intensity and TC genesis performance are computed and evaluated based on the ATCF best track files which contain storm position and intensity information at

6-hourly intervals (Miller et al. 1990; Sampson and Schrader 2000).

#### a. Track forecasts

Figure 2 shows best tracks of the 105 observed TCs and the corresponding forecast tracks from GFS, FV3\_zc, and FV3\_mp. In these plots, tracks of model forecast are shown only every 5 days for each case to avoid excessive clustering graphics on the map. The GFS and both versions of the fvGFS demonstrate reasonable TC forecast tracks. The differences among the three sets of forecast tracks are very subtle.

To further quantify forecast performance of TC track in all models, TC track forecast errors were computed based on the best track data at every 12-h interval. Homogeneous comparisons of basinwide mean TC track forecast errors along with the forecast lead time are shown in Fig. 3. In the North Atlantic Ocean (Fig. 3a), the GFS and the two fvGFS configurations show similar track errors in the first 48 h. However, both FV3\_zc and FV3\_mp show improved track forecasts compared to the GFS after the 60-h lead time. At the 72-h lead time, FV3\_zc and FV3\_mp have 16.2% and 18.7% track error reductions, respectively, compared to the GFS. Similar results can be found in the northwest Pacific basin, where the TC track forecast errors of the two fvGFS forecasts are smaller than those of GFS after 48-h model lead time (Fig. 3d), with 9.6% and 11.4% error reductions, respectively, at 120 h.

In the other two North Pacific basins, the track forecast errors of GFS and fvGFS are relatively close. Results indicate slightly lower errors for FV3\_mp after 84-h lead time than GFS in the northeast Pacific basin (Fig. 3b). Although FV3\_zc shows the largest errors among the three models after 144 h, the differences of FV3\_zc to the other two runs are less than 50 km. In the north-central Pacific, the three runs show very similar TC track forecast errors, but FV3\_zc and FV3\_mp are generally better than GFS at most lead times after 60 h (Fig. 3c).

The results are different in the north Indian Ocean and the Southern Ocean. The GFS shows better TC track forecasts than fvGFS runs in these two regions (Figs. 3e,f). Based on the 95% confidence levels shown in Fig. 3, none of the forecast shows statistically significantly better or worse results in any of the regions. However, the case numbers are quite small in the north Indian Ocean (Fig. 3e), which leads to a lower significance of the result in this area compared to other basins. We also note that TC track errors in the Southern Ocean are larger than those in the Northern Hemisphere oceans (cf. Figs. 3f and 3a–e), especially for FV3\_zc and FV3\_mp forecasts. It may indicate some deficiencies exist in the fvGFS forecasts that are related to the large-scale steering flow in the mixing of seasons in the Southern Hemisphere.

Comparing the two fvGFS forecasts, it was found that FV3\_zc generally showed smaller TC track errors than FV3\_mp in the north Indian Ocean and in the Southern Ocean (Figs. 3e,f). In the North Atlantic basin, FV3\_mp showed slight improvements at 36–60-h lead time (Fig. 3a). To better present the results in the northern Pacific Ocean, Fig. 4a shows the mean TC track errors based on all cases in the three Pacific basins. The improvements of FV3\_mp compared to FV3\_zc are apparent after the 120-h lead time. The largest difference is shown at the 132-h lead time, with a 6.5% reduction of error by FV3\_mp compared to FV3\_zc.

An overview of TC track forecasts from the three models can be shown by presenting the global mean errors (Fig. 4b). The GFS produced slightly larger errors than the forecasts from the two fvGFS versions up to the 168-h lead time. For the two fvGFS track forecasts, FV3\_zc performed slightly better than FV3\_mp after the 84-h lead time. This indicates that the impact to the TC track forecast from updating the FV3 dynamical core in the GFS is positive. However, the updated version of fvGFS with the GFDL microphysics scheme produced slightly degraded the performance of TC track forecasts.

#### b. Intensity forecasts

To investigate the performance of TC intensity forecasts, the wind–pressure relationships of TCs in GFS and FV3\_zc are compared to the best track data in the six regions in Fig. 5. The maximum 10-m wind speed ( $\text{m s}^{-1}$ ) and minimum sea level pressure (SLP, hPa) of the model storms used here are based on the tracker. For intense cyclones with observed intensities exceeding  $40 \text{ m s}^{-1}$ , there is clearly a much better relationship between SLP and maximum 10-m wind speed for FV3\_zc than for the GFS. The model configuration of FV3\_zc uses a physics package nearly identical to that used in the GFS, while the horizontal resolutions of the two forecasts are also very close. Therefore, we believe that the differences shown in Fig. 5 are primarily from the replacement of the dynamical core. It is a very encouraging result that an advanced dynamical core is contributing to improve pressure–wind relationship for TC in a global model.

However, it was also found that there are more intense hurricanes/typhoons (with maximum wind speed over  $60 \text{ m s}^{-1}$ ) in FV3\_zc than observed in the northeast and north-central Pacific basins (Figs. 5b,c). In the North Atlantic, for category 1 hurricanes and above (with maximum wind speed stronger than  $33 \text{ m s}^{-1}$ ), the wind speeds of FV3\_zc are slightly weaker than observed, despite the fact that the number of cyclones with minimum SLPs in the deepest range ( $<930 \text{ hPa}$ ) are overpredicted

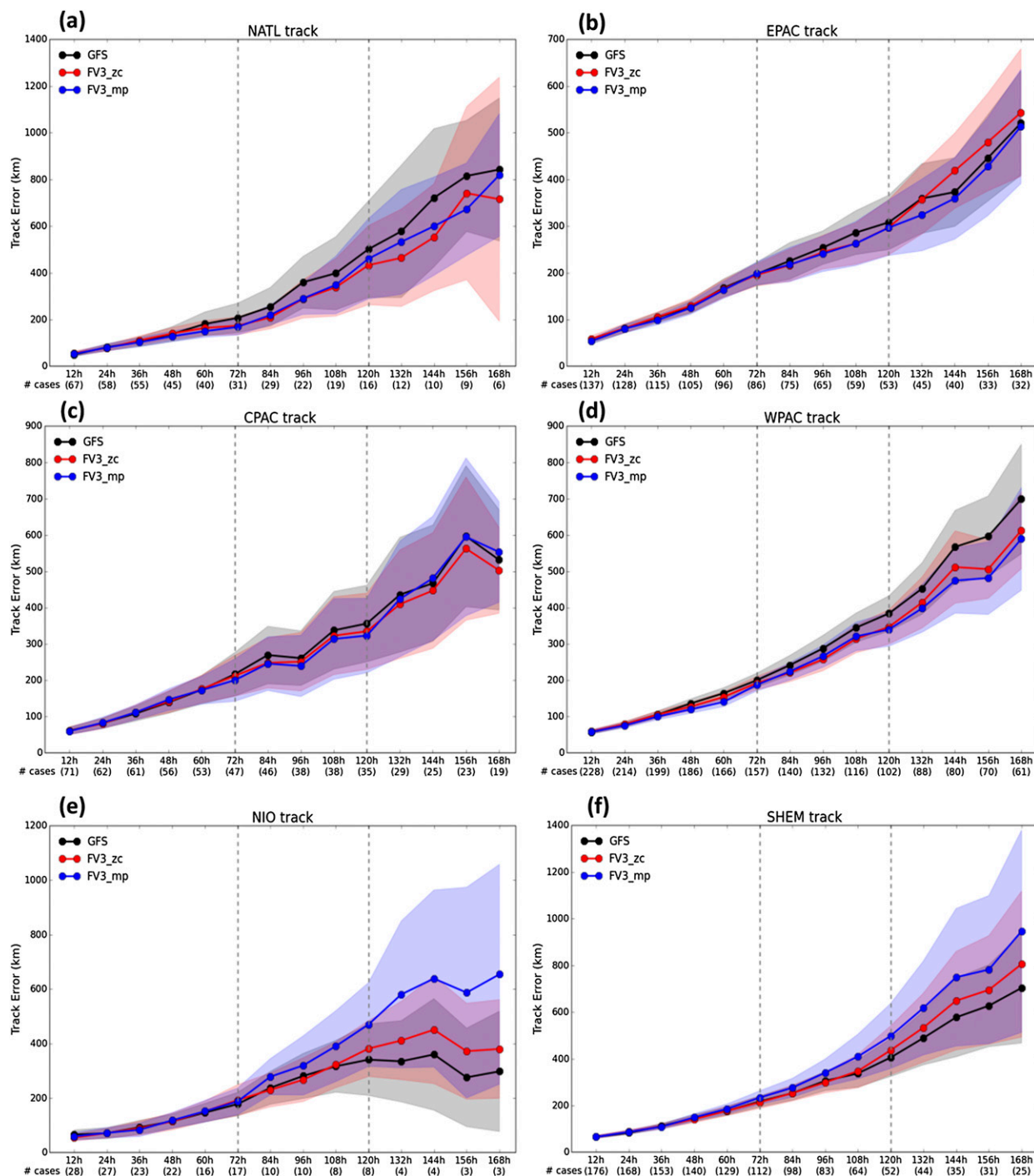


FIG. 3. Basinwide mean TC track forecast errors (km) along with the model forecast lead time for GFS (black), FV3\_zc (red), and FV3\_mp (blue) in (a) the North Atlantic basin, (b) the northeast Pacific basin, (c) the north-central Pacific basin, (d) the northwest Pacific basin, (e) the north Indian Ocean, and (f) the Southern Ocean. The 95% confidence levels for each model are indicated by the same transparent color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom of each abscissa. Vertical gray dotted lines are indicated 72 and 120 h.

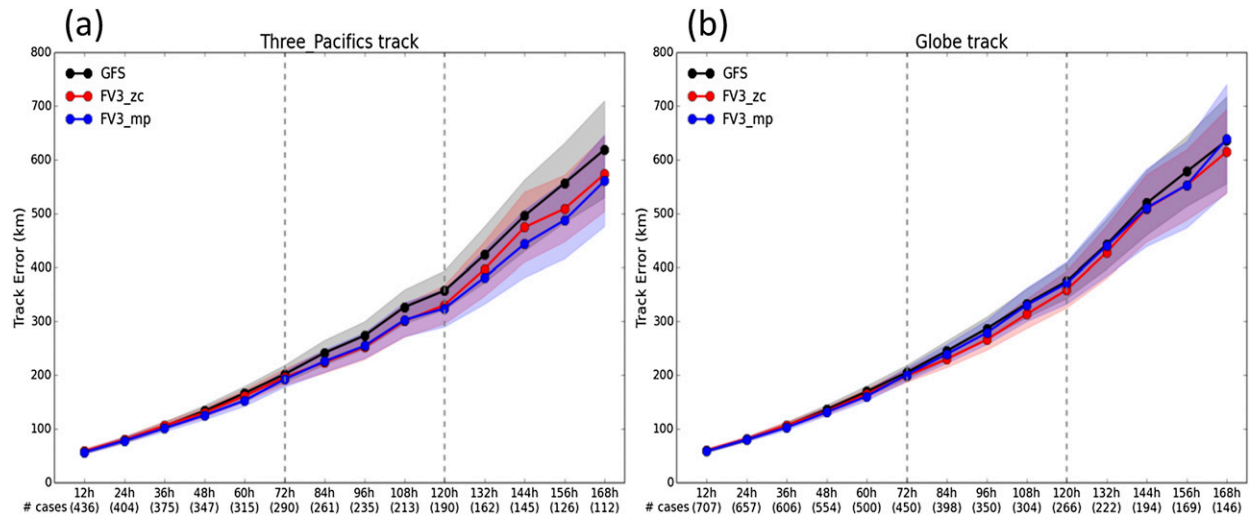


FIG. 4. Mean TC track forecast errors for GFS (black), FV3\_zc (red), and FV3\_mp (blue) in (a) the three Pacific basins and (b) the globe along with the model forecast lead time. The 95% confidence levels for each model are indicated by the same transparent color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom of each abscissa. Vertical gray dotted lines are indicated at 72 and 120 h.

(Fig. 5a). This overprediction of central pressure but underprediction of 10-m winds is also found in the northeast Pacific basin for storms that have maximum wind speeds of  $40\text{--}60\text{ m s}^{-1}$  and minimum SLPs of  $960\text{--}920\text{ hPa}$  (Fig. 5b), as well as in the northwest Pacific basin and the Southern Ocean for storms with maximum wind speed about  $60\text{--}70\text{ m s}^{-1}$  (Figs. 5d,f).

The above overpredicted intensities of strong storms shown in FV3\_zc can be improved in FV3\_mp. Comparing Fig. 5c to Fig. 6c, we can see that the large number of overpredicted storm intensity values shown in FV3\_zc is greatly reduced in FV3\_mp in the north-central Pacific basin. The wind–pressure relationship of FV3\_mp is also better matched to the observation in Fig. 6c. However, in the northeast Pacific, the northwest Pacific and the north Indian Ocean, the number of most intense storms in FV3\_mp are fewer compared to the observations (Figs. 6b,d,e). Nonetheless, the wind–pressure relationships in these areas are still quite close to the observations.

The results shown in Figs. 3, 5, and 6 for FV3\_zc and FV3\_mp indicate that the main benefit of replacing the Zhao–Carr scheme with the GFDL microphysics scheme is the improvement in intensity prediction rather than track prediction. This result is further confirmed by analyzing global mean forecast errors of track and intensity versus lead time (Figs. 4b and 7). Figure 7a shows that the differences in the absolute error of maximum 10-m wind among the three sets of forecasts are not significant. During the first 84 h, the two fvGFS versions show smaller errors than those of the GFS, while the FV3\_mp is the best of the three. After the 120-h lead time, GFS has lower errors than FV3\_ZC, but the errors

of FV3\_mp remain smaller than those of GFS. The biases of maximum 10-m wind speed shown in Fig. 7c indicate that the GFS underpredicted TC intensities (negative intensity bias) through 132 h, while FV3\_zc overpredicted intensities (positive intensity bias) during the entire 7-day forecast. In contrast, the FV3\_mp shows very small negative intensity biases compared to FV3\_zc. Noted that at 12-h lead time, the GFS has a large negative bias about  $3.5\text{ m s}^{-1}$ , but there is almost no bias in either of the fvGFS forecasts that start from the same GFS initial condition. It may indicate that fvGFS can simulate the marginal resolved (13 km) TC structures better than the GFS.

The improvement of intensity forecasts in the FV3\_mp is also demonstrated by examining the absolute error and bias of the minimum SLP (Figs. 7b,d). Based on this measure, the FV3\_mp shows the best intensity forecast performance of the three models (Fig. 7b). Note that FV3\_mp shows significantly smaller errors after 48 h compared to the GFS. Similar to its result for the maximum 10-m wind, the FV3\_mp shows only a slight bias during the 7-day forecasts (Fig. 7d), while both GFS and FV3\_zc show a positive intensity bias with an increasing trend toward deeper pressures with increasing lead time. In summary, these results demonstrate the potential positive impact in improved prediction of TC intensity in the fvGFS by upgrading the microphysics with the more advanced GFDL microphysics scheme.

#### 4. TC genesis forecasts in fvGFS

Unlike verifying TC track and intensity forecasts, investigating the model performance for TC genesis requires



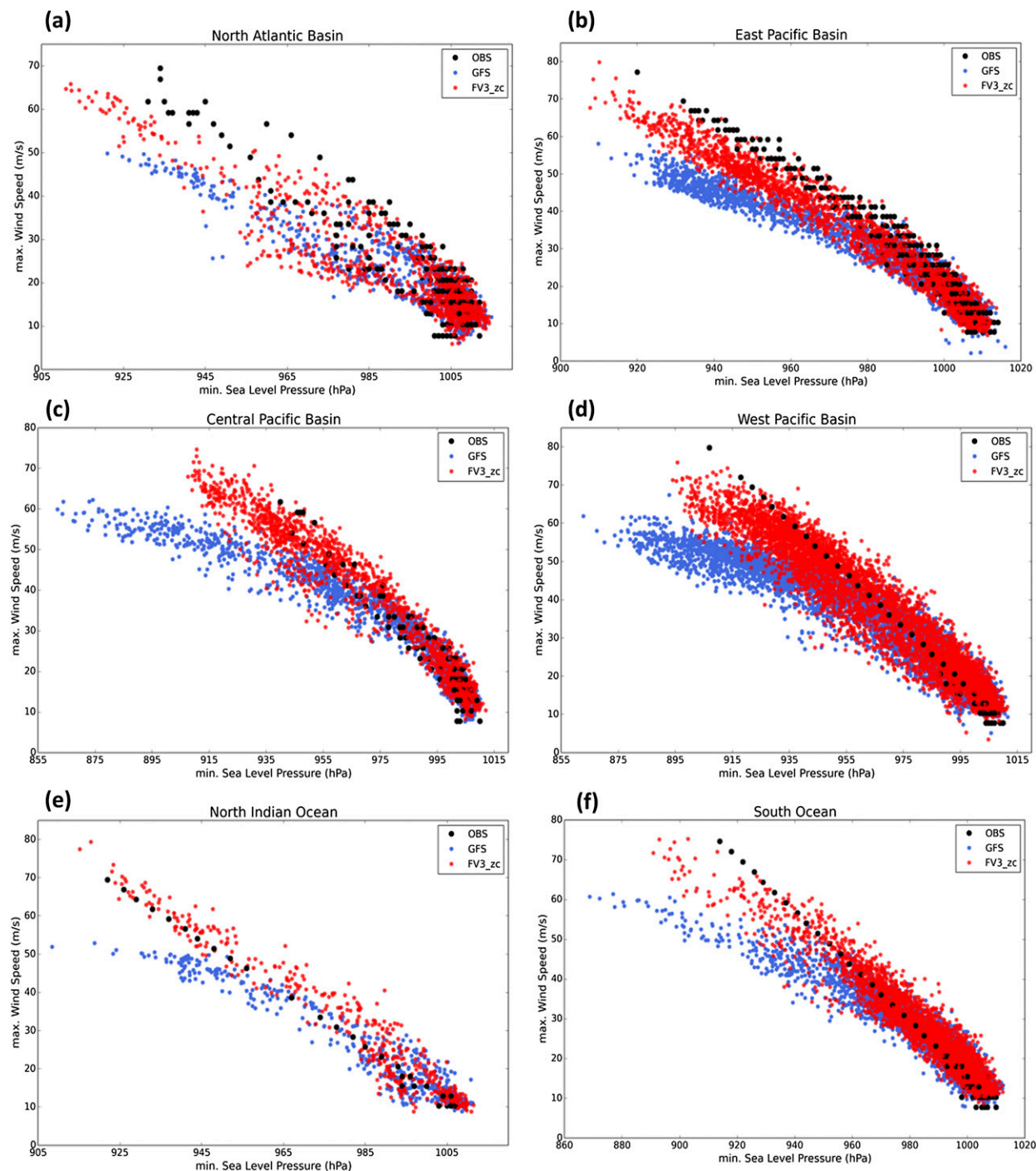


FIG. 5. The relationship of maximum 10-m wind ( $m s^{-1}$ ) and minimum sea level pressure (hPa) for TCs in (a) the North Atlantic Ocean, (b) the northeast Pacific basin, (c) the north-central Pacific basin, (d) the northwest Pacific basin, (e) the north Indian Ocean, and (f) the Southern Ocean. Forecast data are plotted every 6 h of lead time during the 168 h forecast. The observations every 6 h from ATCF best track data are denoted in black dots. Forecasts of GFS cyclones are in blue dots, and of FV3\_zc are in red.

considering the 10-day forecasts which were initialized before the observed TC genesis time. However, the periods of these 10-day forecasts should cover the observed TC genesis. The observed TC genesis times are

based on the first tropical depression (TD) record in the ATCF best track data. After excluding TCs that exist in the initial conditions, all TCs found by the GFDL simple tracker in these 10-day forecasts were counted as genesis

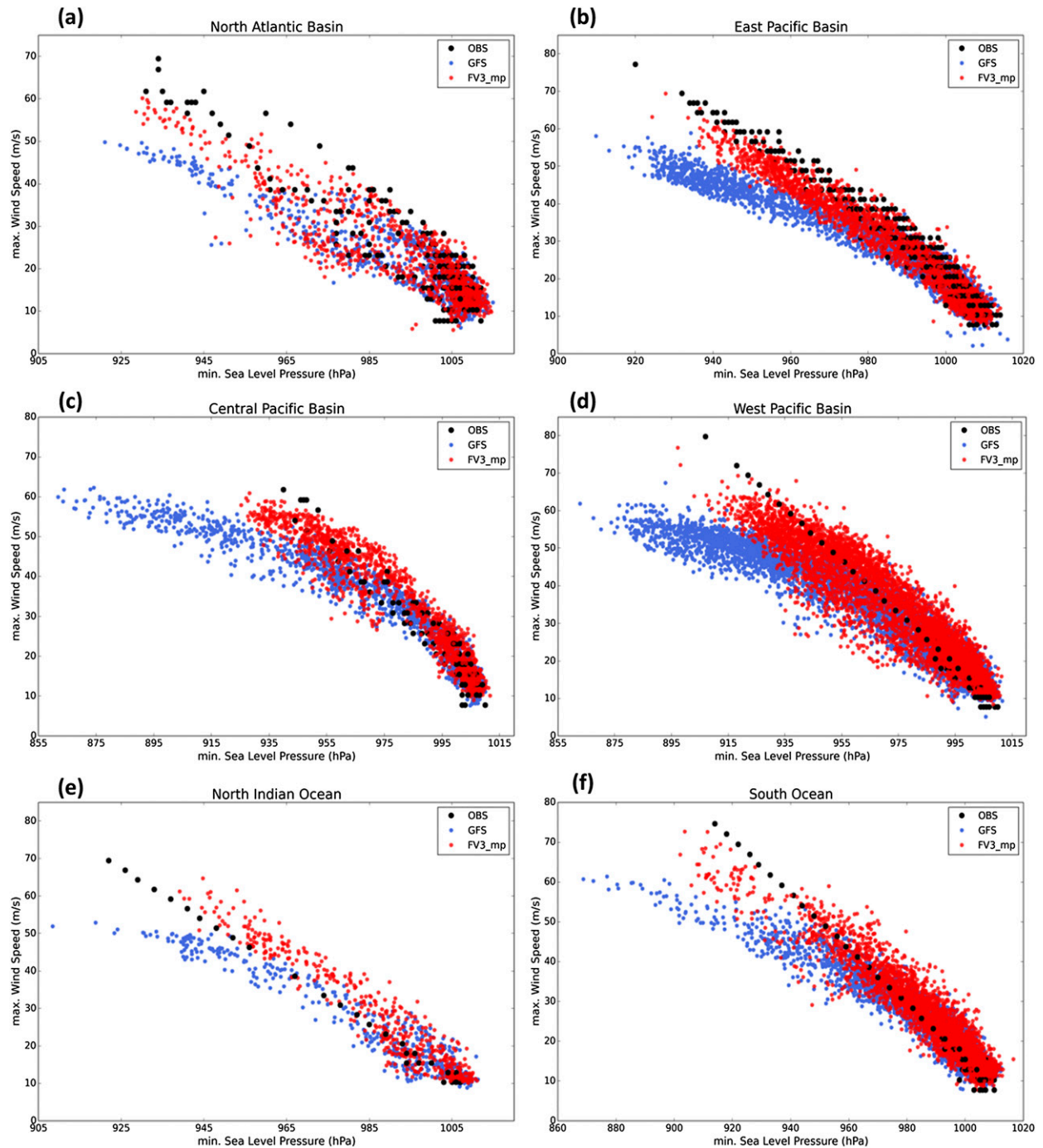


FIG. 6. As in Fig. 5, but red dots are for FV3\_mp.

events. If the model genesis storm showed a matched track to the observed TC, it is a “hit event.” Figure 8 illustrates all related timelines of a sample hit event for a given model. The storm genesis of a hit event occurred during a 10-day model forecast (yellow line), which is initialized prior to the observed TC genesis. The life cycle of an observed TC is denoted by black hurricane

symbols, with the filled symbol indicating the observed genesis time. If genesis occurs in the model for the forecast of this observed TC at any lead time, then the blue line indicates the observed genesis lead time (OLT) for this case. The OLT is a nominal lead time value that indicates the difference in time between the forecast initial time and the time at which observed genesis

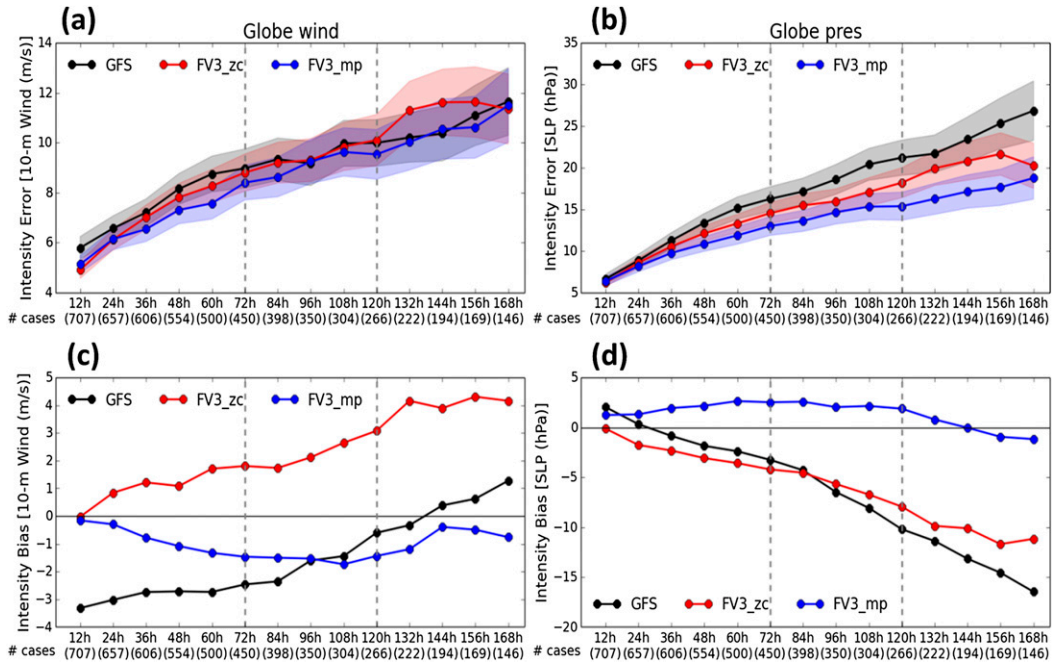


FIG. 7. Global mean TC intensity errors and biases. (a) Absolute error of the maximum 10-m wind speed ( $m s^{-1}$ ) along with the model forecast lead time for GFS (black), FV3\_zc (red), and FV3\_mp (blue). Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom of each abscissa. (b) As in (a), but for minimum sea level pressure (hPa). (c) As in (a), but for the bias of the maximum 10-m wind speed ( $m s^{-1}$ ). (d) As in (b), but for the bias of the minimum sea level pressure (hPa). The 95% confidence levels for each model are indicated by the same transparent color shading in (a),(b). Vertical gray dotted lines are indicated at 72 and 120 h.

occurs. A potential range of solutions for the lead time at which the actual forecast genesis may occur for this sample model forecast is indicated by the red lines. For this model forecast, TC genesis could occur at the same time as the observed TC genesis, or before or after the observed TC genesis (these three scenarios are represented by the three red hurricane symbols). This time span from the model initial time to the model-predicted genesis lead time is referred to as the model genesis lead time (MLT). The differences between the MLT and OLT (DMO) are illustrated by the green arrows, and the lengths of these DMOs indicate how accurate a

model is in generating storms at the observed genesis time.

To consider a model storm as being the same as an observed storm, the maximum time difference of model genesis and observed genesis is set to 72 h. The distance of model and observed storms needs to be within a  $5^{\circ}$  by  $5^{\circ}$  box centered at the observed storm at their first matched time. Model genesis storms which do not have tracks that match to observed TCs are categorized as “false alarms.” The sum of hit events and false alarms is equal to the number of total genesis events in the model. The forecasts that are initialized within the 10 days

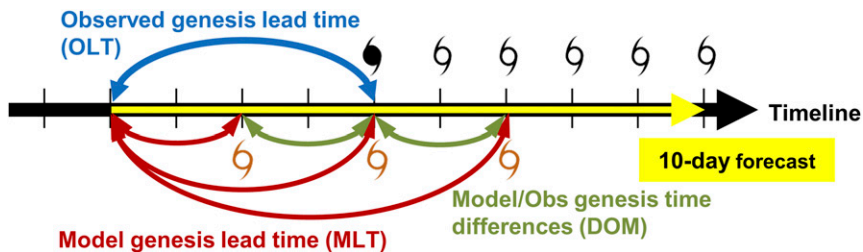


FIG. 8. Illustration of timelines for hit events of storm genesis. The timeline increments show the 6-h interval in the best track data and model forecasts. Detail descriptions can be found in the first paragraph of section 4.

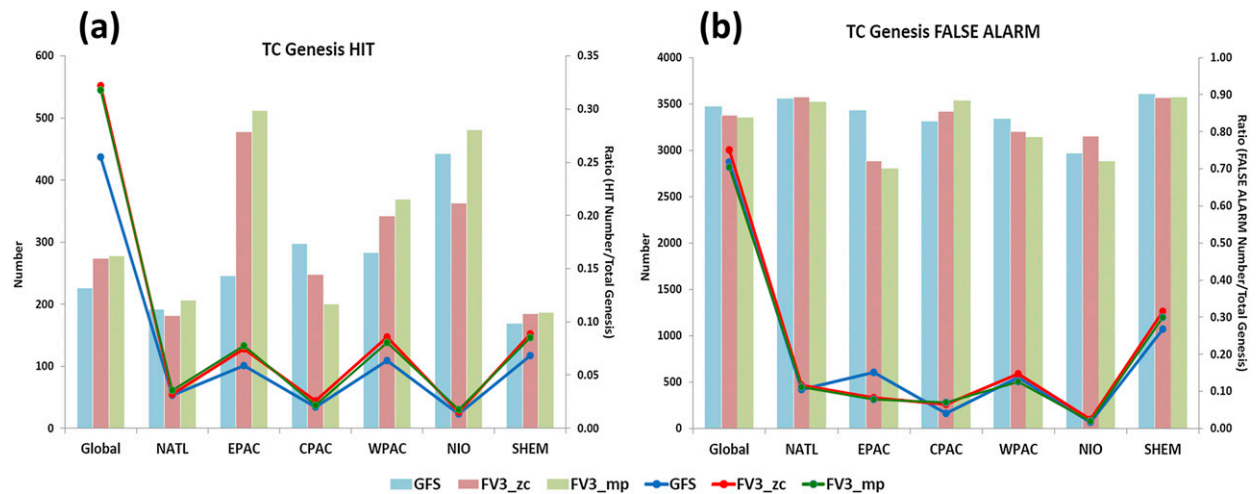


FIG. 9. (a) Numbers of total genesis hit events across all forecasts and all lead times (solid lines; ordinate to the left) and ratios of hit events to the total number of genesis forecast events (bars; ordinate to the right) of GFS (blue), FV3\_zc (red), and FV3\_mp (green) in the globe and six basins. Abbreviations used for the abscissa are the same as in Table 2. (b) As in (a), but for numbers of false alarms and the ratio of false alarms to the total number of genesis forecast events.

before the observed TC genesis time but do not predict the genesis event of a given observed storm are considered missing cases.

#### a. Analyses of hit events, false alarms, and missing cases

Considering all forecasts at all lead times, Fig. 9 shows the number of general hit events and false alarms, as well as their ratios of model hits to total observed genesis events, over the globe and in the 6 individual basins for GFS, FV3\_zc, and FV3\_mp. Both FV3\_zc and FV3\_mp show more hit events and higher hit ratios than GFS over the globe (Fig. 9a). The FV3\_mp shows the highest hit ratio in the five regions but not in the north-central Pacific basin, where the GFS performs the best. For the North Atlantic basin and the north Indian Ocean, FV3\_zc shows similar hit numbers but lower hit ratios than GFS. Note that in the northeast Pacific basin, the hit ratio of FV3\_mp is 2 times larger than that of GFS, which can be explained later by the false alarm results.

The numbers of both hit events and false alarms are modulated and partly determined by the values of the thresholds applied on the sorter in the simple storm tracker package (e.g., the length of storm lifetime). In addition, false alarms are more sensitive to the settings of thresholds than hit events since their case numbers are much larger than those of hit events in the model. Therefore, the relative differences of numbers and ratios among the three sets of model forecasts are more important than their absolute values. In Fig. 9b, large differences of false alarms between the GFS and fvGFS are shown in the northeast Pacific basin, where the GFS

forecasts show more false alarms in this basin than forecasts from the two fvGFS versions. This explains the complementary performance in hit ratio for the two fvGFS versions found in Fig. 9a for the northeast Pacific basin.

Figure 10 shows monthly numbers of total genesis events (lines) and ratios of hit events to total genesis events (bars) of the three sets of forecasts in the six basins. In the North Atlantic basin, all three sets of forecasts show that the numbers of monthly total genesis forecast events gradually increase from late boreal spring into summer and reaching the peak in September (Fig. 10a), which is consistent with observations of TC seasonal development in this basin. Similar variations can be found in the northwest Pacific basin (Fig. 10d). In these two basins, the total genesis forecast numbers in the peak of the TC season are about 2–3 times larger than those in the off-peak months of the year. Also, differences of their monthly total genesis numbers between the GFS and fvGFS forecasts are relatively small.

In contrast to the above two basins, in the northeast Pacific basin, there are substantial variations between the GFS and fvGFS forecasts during the peak months (Fig. 10b). For the GFS, there are about 150 genesis forecast events in the peak month of July, which is more than its peak number in the northwest Pacific basin. In contrast, the number of observed TC cases in the northeast Pacific basin is only half of those in the northwest Pacific basin in July. These overpredicted genesis events cause large numbers of false alarms as well as the low hit ratios of the GFS in the northeast Pacific TC season. Large improvements of the number



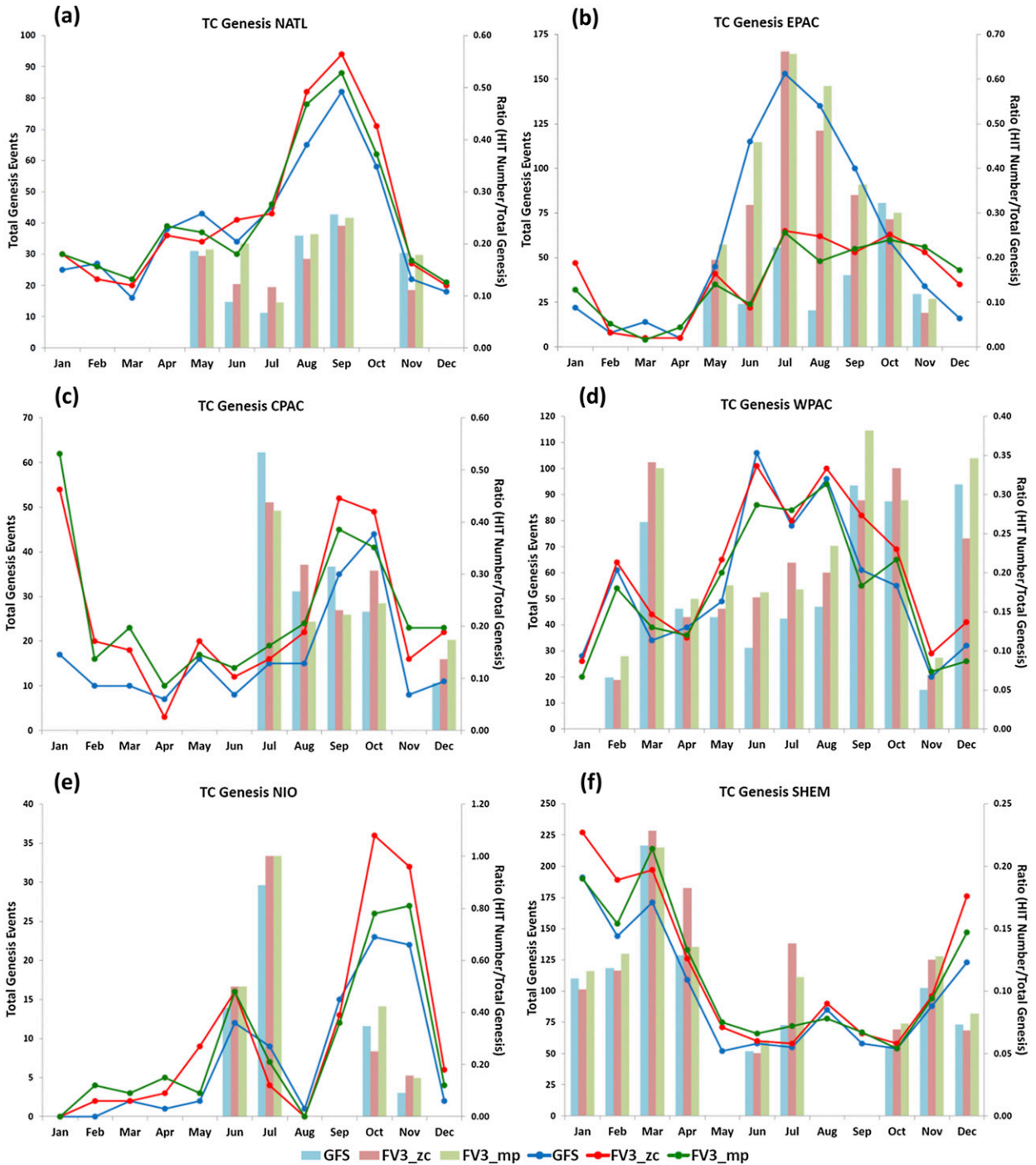


FIG. 10. Monthly numbers of total genesis forecast events (solid lines; ordinate to the left) and ratios of hit events to the total genesis forecast events (bars; ordinate to the right) of GFS (blue), FV3\_zc (red), and FV3\_mp (green) in (a) the North Atlantic basin, (b) the northeast Pacific basin, (c) the north-central Pacific basin, (d) the northwest Pacific basin, (e) the north Indian Ocean, and (f) the Southern Ocean.

of total genesis forecast events and hit ratios are shown in the two fvGFS forecasts. However, in the north-central Pacific basin (Fig. 10c), both fvGFS versions show many genesis forecast events (false alarms) in January and

December that are not seen in the observations. These off-season false alarms causes the relative low hit ratios of FV3\_zc and FV3\_mp in the north-central Pacific basin shown in Fig. 9. The two fvGFS forecast also

TABLE 2. Numbers of observed TCs generated during 16 Jan 2015 to 16 Jan 2016 and numbers of missed TCs in the genesis forecasts of GFS, FV3\_zc, and FV3\_mp in the North Atlantic basin (NATL), the northeast Pacific basin (EPAC), the north-central Pacific basin (CPAC), the northwest Pacific basin (WPAC), the north Indian Ocean (NIO), and the Southern Ocean (SHEM).

Basins	NATL	EPAC	CPAC	WPAC	NIO	SHEM
No. of observed TCs	12	22	9	28	5	26
No. of missed TCs in the genesis forecasts of						
GFS	0	2	1	2	0	1
FV3_zc	1	0	1	1	0	0
FV3_mp	1	0	0	1	0	0

demonstrated more false alarms than the GFS in January and December in the northeast Pacific basin. It is speculated that the applied sorter is not as effective at filtering extratropical storms in these two basins as in other regions. However, the impact on the GFS forecasts appears relatively minor which will need further investigation. In the north Indian Ocean and the Southern Ocean (Figs. 10e,f), the two fvGFS forecasts show generally higher hit ratios than the GFS forecast. Note that the GFS usually shows fewer total genesis numbers than the fvGFS, which results in slightly fewer false alarms in the off-peak months (e.g., February to May in the north Indian Ocean; May and September in the Southern Ocean).

Neither the GFS nor the fvGFS predicted all observed TCs' genesis during the 16 January 2015–16 January 2016 period. Table 2 lists the numbers of observed TCs and numbers of missed TCs for the genesis forecasts of GFS, FV3\_zc, and FV3\_mp in the six basins. Typhoon Mekkhala in the northwest Pacific basin and Cyclone Bansi and TS Chedza in the Southern Ocean are not included in Table 2 since they formed earlier than 16 January 2015. The GFS missed 6 of 102 TC's genesis during the focus period. Five of the missed observed TCs were in the northern Pacific Ocean. The global numbers of missed observed TCs dropped to 3 and 2 in the FV3\_zc and FV3\_mp forecasts, respectively. Both fvGFS forecasts showed improvement in the three North Pacific basins, but missed 1 TC genesis in the North Atlantic basin.

In addition to comparing the number of completely missed observed TCs' genesis, the ratios of missing storms in each of the three sets of forecasts in each basin are listed in Table 3. The miss ratio can be computed as the ratio of number of missing cases to the number of expected genesis observations, which is 9 or 10 for each observed TC (one for each day of lead time) except for 4 for TS Niko in the Southern Ocean. For TS Niko, there were only 4 forecasts in our model dataset before it formed at 0600 UTC 19 January 2015. The GFS shows the highest miss ratio in all basins, which is consistent with the results shown in Table 2. The FV3\_zc produced lower miss ratios compared to the GFS and FV3\_mp in the north-central Pacific basin, the northwest Pacific

basin, and the Southern Ocean. The FV3\_mp shows the lowest miss ratio among the three sets of forecasts in the other three basins. All forecasts show higher miss ratios in the north-central Pacific basin compared to the other basins. As shown in Table 2, both the GFS and FV3\_zc missed 1 of 9 TC geneses in the north-central Pacific basin, an area with a relative high frequency of missing events.

Generally speaking, FV3\_mp shows the best performance of the three models in the above hit events (highest ratio of hit events in 5 of 6 basins), false alarms (lowest ratio of false alarms in 5 of 6 basins), and missing cases comparisons (only missed 2 of 106 TC's genesis, but 3 and 6 for FV3\_zc and GFS, respectively), followed by FV3\_zc. The GFS forecast only shows the highest/lowest ratios of his events/false alarms in the north-central Pacific. The performance of FV3\_mp demonstrates that the modern dynamical core with the upgraded microphysics scheme can improve TC genesis forecasts.

#### b. Analyses of the forecast lead time of hit events

Another way to investigate model storm genesis is through the lead time lengths of hit events, which is an innovation beyond previous studies. In this section, the OLT, MLT, and their differences (DMO) (please refer Fig. 8 for the definitions of different lead time lengths) are compared for the three set of forecasts. If a model shows longer OLT, that means it can correctly predict TC genesis earlier than other models. For an observed TC, there could be many hit events that happened in successive 10-day model forecasts, which were initialized earlier than the observed genesis time. When comparing the lengths of OLT in different models, we

TABLE 3. Miss ratios (%) in the genesis forecasts of GFS, FV3\_zc, and FV3\_mp in the six basins (abbreviations are the same as in Table 2). The lowest miss ratios of the three sets of forecasts in each basin are marked in bold.

Basins	NATL	EPAC	CPAC	WPAC	NIO	SHEM
GFS	55.8	53.2	62.2	60.5	53.1	53.0
FV3_zc	54.2	40.7	<b>51.1</b>	<b>46.7</b>	46.9	<b>39.0</b>
FV3_mp	<b>49.2</b>	<b>38.4</b>	58.9	50.0	<b>38.8</b>	41.4

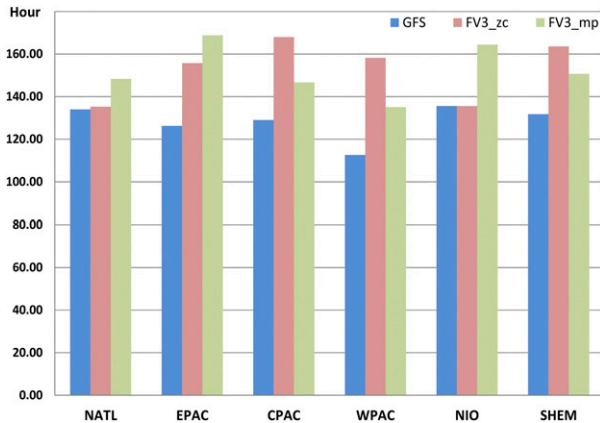


FIG. 11. Mean values of maximum observed genesis lead time of all storms in each basin of GFS (blue), FV3\_zc (red), and FV3\_mp (green). Abbreviations of the six basins used for the abscissa are the same as in Table 2.

only consider the maximum (the longest) OLT of each observed TC.

Figure 11 shows the mean values of the maximum OLT of all observed TCs in each area for GFS, FV3\_zc, and FV3\_mp. For TCs in the North Atlantic basin, the mean maximum OLTs are at 134, 135, and 148 h for the GFS, FV3\_zc, and FV3\_mp, respectively. This means that the average earliest time for the GFS to successfully predict a hit event is 134 h before the observed TC genesis, while it is 1 h earlier for FV3\_zc to predict a hit event in this basin. Meanwhile, FV3\_mp can predict a hit event 14 h earlier than GFS. In other basins, both fvGFS forecasts also show longer OLTs than GFS except for the north Indian Ocean where the GFS and FV3\_ZC have about the same length of OLTs. The longest OLTs (168 h) are achieved by FV3\_mp in the northeast Pacific basin and by FV3\_zc in the north-central Pacific basin. The Mann–Whitney *U* test (also called the Wilcoxon rank-sum test) is applied to test the significance between the OLTs from any of the two forecasts in each basin. The statistics (*U*) and probability value (*p* value) are listed in Table 4. The difference between the GFS and the FV3\_mp in the northeast Pacific basin and the difference between the GFS and the FV3\_zc in the northwest Pacific basin are statistically significant (*p* value < 0.05, two tailed).

The differences between MLT and OLT are DMO (green arrows in Fig. 8). The length of these DMOs indicates how accurate a model is in generating storms at the observed TC genesis time. If a model forecast TC generates exactly on the observed TC genesis time, the DMO of this hit event is “zero.” A positive (negative) DMO means that the hit event occurs later (earlier) than the observed TC genesis time. According to our definition

TABLE 4. The statistics (*U*) and probability values (*p* values) of the Mann–Whitney *U* test (also called the Wilcoxon rank-sum test) between the OLTs from any of the two forecasts the six basins (abbreviations are the same as in Table 2). Bold numbers indicate that the OLT differences of the two forecasts are statistically significant (*p* value < 0.05, two tailed).

Statistics	GFS vs FV3_zc		GFS vs FV3_mp		FV3_zc vs FV3_mp	
	<i>U</i>	<i>p</i> value	<i>U</i>	<i>p</i> value	<i>U</i>	<i>p</i> value
NATL	67.5	0.951	59.5	0.711	51.0	0.548
EPAC	159.0	0.127	138.5	<b>0.041</b>	203.5	0.372
CPAC	16.0	0.101	31.0	0.664	46.5	0.334
WPAC	172.5	<b>0.002</b>	258.5	0.101	475.5	0.556
NIO	12.5	0.917	8.5	0.463	11.0	0.834
SHEM	232.5	0.082	267.0	0.278	373.5	0.521

of hit events, the maximum length of DMO is 72 h. Figure 12 shows the fractions of global total hit events occurring within a certain DMO length for each model forecast. For 0-h DMO hit events, FV3\_zc and FV3\_mp show slightly larger fractions than the GFS. The FV3\_zc shows the largest fractions among the three sets of forecasts at 6- and 12-h DMOs. The largest fraction of the hit events in the GFS occurred by the 48-h DMO, but the differences among the three sets of forecasts are smaller than 3%. Note that more than 88% of the hit events in any of the forecasts occur by the 48-h DMO. For all three sets of forecasts, there are more than 60% hit events occurring before the observed TC genesis times rather than after.

The fractions of hit events within different DMO lengths for individual basin are shown next in Fig. 13. A large variation can be found among the six basins. In the North Atlantic and the north-central Pacific basins, FV3\_zc shows the highest fractions, except for hit events within 48-h DMO in the North Atlantic and for those within 0-h DMO in the north-central Pacific basin (Figs. 13a,c). In the northeast Pacific and Southern Ocean, the GFS generally shows the highest fractions among the three (Figs. 13b,f). The FV3\_mp shows generally higher fractions than the GFS and FV3\_zc in the northwest Pacific basin (Fig. 13d). However, in the north Indian Ocean, FV3\_mp shows lowest fractions for the 12-, 24- and 48-h DMOs compared to GFS and FV3\_zc. Meanwhile, the differences between GFS and FV3\_zc are relatively small in this basin (Fig. 13e).

All three sets of forecasts show more hit events occurring before the observed TC genesis time rather than after in all six basins. This is related to the choice of the “observed TC genesis time,” which is defined as the first “tropical depression” record for each TC in the ATCF best track data. However, the model predicted genesis events often occur during the precursor stage of the

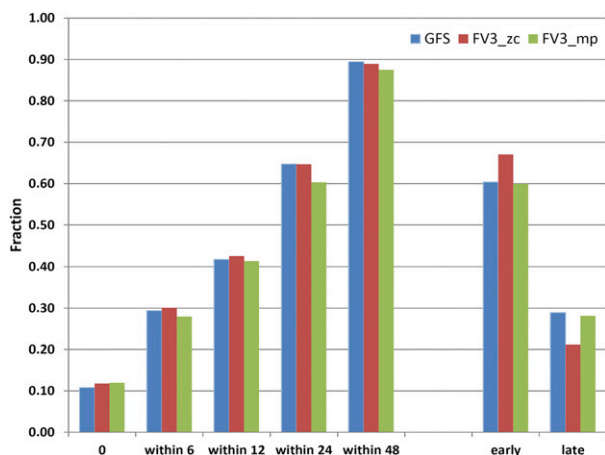


FIG. 12. Fractions of global total hit events occurring within a certain DMO length for GFS (blue), FV3\_zc (red), and FV3\_mp (green). “0” is for hit events which happened at the observed genesis time. “Within 6 (12, 24, or 48)” is for hit events with DMO lengths in 6 (12, 24, or 48) hours. “Early (late)” is for all hit events with negative (positive) DMOs.

observed TC. A test was conducted to instead use the first record in the ATCF best track data as the observed TC genesis time, which could be a disturbance (DB), a low (LO), or a tropical wave (WV). It was found that a substantial number of early events shifted to later times in most basins (the North Atlantic, the north-central Pacific, the northwest Pacific basins, and the Southern Ocean; figures not shown). Similar results were found for all three sets of forecasts. Those shifted hit cases were generated during the precursor stage of the observed TCs. It is interesting that the ratios of “early” and “late” events are not changed too much in the northeast Pacific basin and the north Indian basin (figures not shown). This is because the periods of precursor stage, for example, from the first recorded “DB” (“LO” or “WV”) to the first recorded “TD” in the best track data, are shorter in these two basins than those in the other four basins. Therefore, the impacts of using early observed TC genesis times are relatively minor in these two basins.

Finally, mean biases and root-mean-square errors (RMSEs) of the storm genesis intensity (10-m wind speed) in the three sets of forecasts were evaluated in Fig. 14. Hit events with DMO lengths longer than 48 h were excluded to more precisely compare the intensities of model genesis storms to the observations. For hit events occurring at the observed genesis times or within the 48-h DMO before the observed genesis time, the model TC intensities taken at the observed genesis time were compared to the genesis intensities of observed TCs. For hit events occurring within the 48-h DMO after the observed genesis times, the model TC genesis

intensities were compared to the observed TC intensities at the model TC genesis times. All three sets of forecasts under predict their storm genesis intensities in the North Atlantic basin (Fig. 14). However, the two sets of fvGFS forecasts show smaller negative biases and smaller RMSEs compared to the GFS in this basin. A similar underprediction of storm genesis intensities in the GFS forecasts is also apparent in the northeast Pacific, the north-central Pacific basins, and the north Indian Ocean. In the northeast Pacific basin, the large negative bias shown in the GFS forecasts is greatly improved in the FV3\_zc and FV3\_mp forecasts, while the FV3\_mp also shows the smallest RMSEs among the three sets of forecasts. The GFS predicts the best storm genesis intensities in the northwest Pacific basin and in the Southern Ocean, with smallest biases and RMSEs among the three sets of forecasts. Generally, the two fvGFS versions show larger positive bias than the GFS globally. This overpredicted TC genesis intensity in the fvGFS is consistent with the results of the wind–pressure relationship for all TCs shown in Figs. 5 and 6. The overpredictions of TC genesis intensities also cause the relatively large RMSEs in the fvGFS forecasts.

## 5. Summary and discussion

Global models in most operational centers (e.g., NCEP/GFS, ECMWF/IFS, and the Met Office UM), have demonstrated considerable skill at predicting TC tracks in recent years. However, the forecast of TC intensity has remained a challenge for global models. The nonhydrostatic version of the GFDL Finite-Volume Cubed-Sphere dynamical core (FV3) demonstrated considerable accuracy, scalability, and computational efficiency during the NGGPS Dynamical Core Evaluation in 2014–17. This newly selected “engine” for the next generation GFS model is expected to provide a great opportunity for the unification of weather and climate prediction systems, as well as to improve the TC intensity forecasts in the new GFS model.

A new global model, called fvGFS, using the FV3 dynamical core coupled to the physics package from NCEP/GFS was built at GFDL. In this study, we compared TC track, intensity, and genesis forecasts in fvGFS to those in the operational GFS. Two configurations of fvGFS were investigated here. One is the NGGPS submission version (FV3\_zc). The other adopts most of the model settings of FV3\_zc but uses upgraded GFDL cloud microphysics with corresponding model adjustments (FV3\_mp). The performance of global TC forecasts of fvGFS and the operational GFS was investigated based on 363 cases of 10-day forecasts in 2015. The chosen



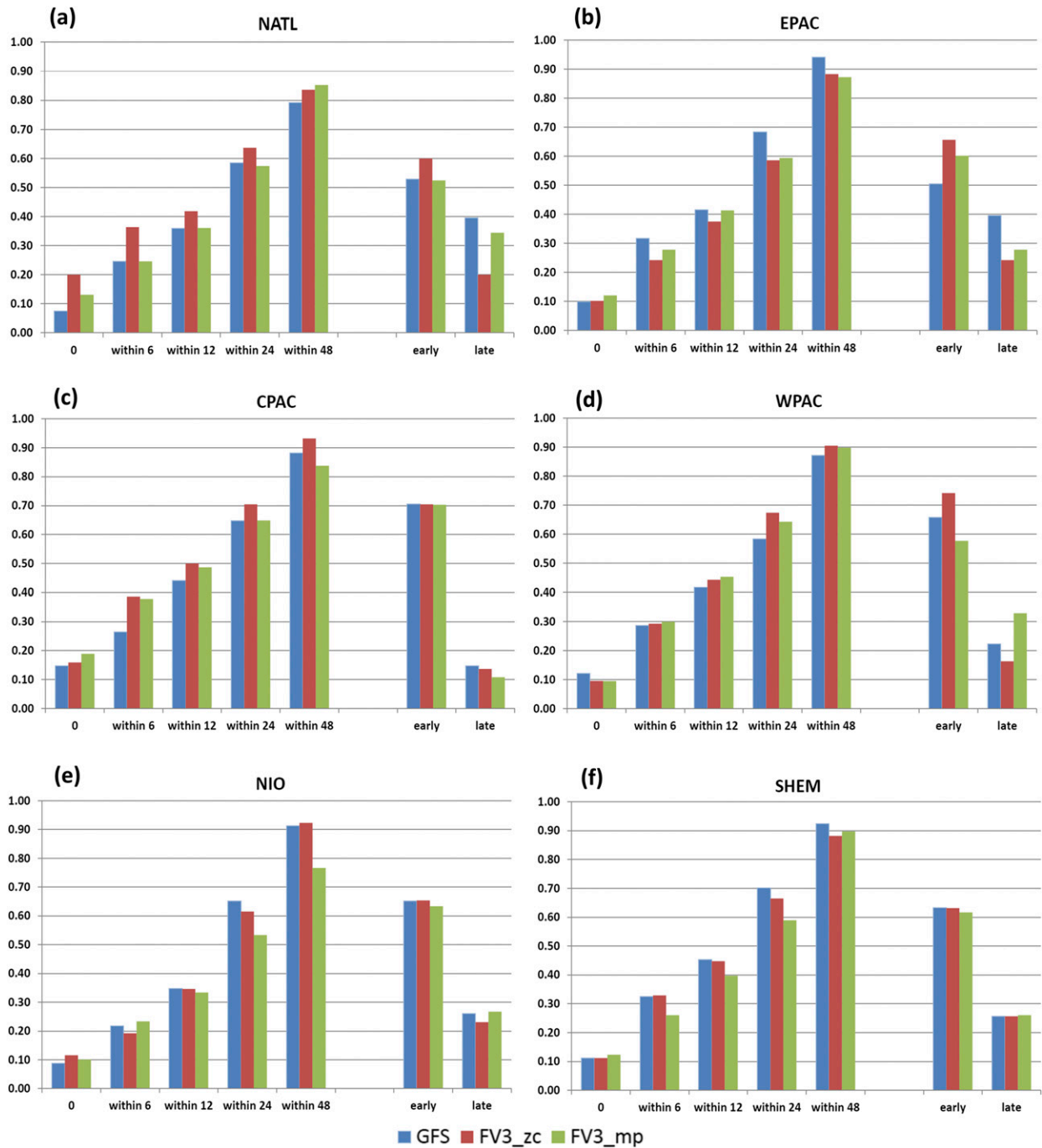


FIG. 13. As in Fig. 12, but for (a) the North Atlantic basin, (b) the northeast Pacific basin, (c) the north-central Pacific basin, (d) the northwest Pacific basin, (e) the north Indian Ocean, and (f) the Southern Ocean.

horizontal resolution of fvGFS is 13 km globally, which is close to that in the operational GFS.

The ATCF data only contains the GFS TC forecast information that was generated from the GFDL TC tracker (Marchok 2002) based on the quarter-degree GFS forecast outputs. To conduct a fair comparison, in

this study, the GFDL simple tracker was applied to the GFS forecasts at its native resolution. TCs in the forecasts from GFS and the two versions of fvGFS were tracked by the same tracker. The errors of TC track and intensity were computed and evaluated based on the ATCF best track data. The use of the GFDL simple

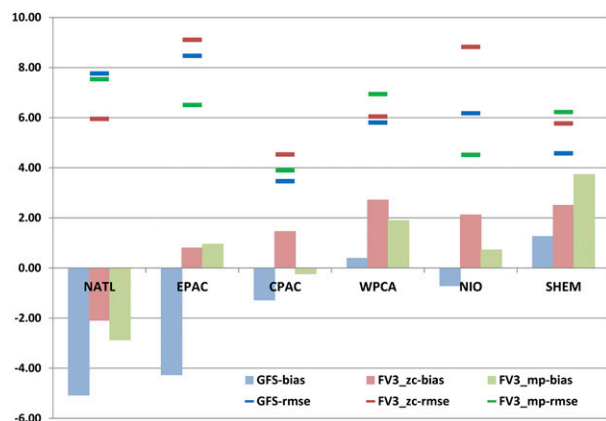


FIG. 14. Mean biases (bars) and root-mean-square errors (horizontal lines) of the storm genesis intensity (10-m wind speed) in GFS (blue), FV3\_zc (red), and FV3\_mp (green). Only hit events occurring within the 48-h DMO are considered. Abbreviations of the six basins used for the abscissa are the same as in Table 2.

tracker also helped the development of the TC genesis verification tools due to its relative simple output format.

Homogeneous comparisons of the basinwide mean TC track forecast errors show that the forecast performance of TC track is better in both versions of fvGFS over the North Atlantic basin and the northwest Pacific Ocean. In the northeast Pacific and the north-central Pacific basins, the track forecast errors of the GFS and fvGFS were mostly comparable. The improvements of TC track forecast of fvGFS relative to the GFS mostly occurred in the Northern Hemisphere. In the Southern Ocean, all model forecasts show larger track forecast errors. It may imply some deficiencies existed in the Southern Hemisphere large-scale forecasts for all models.

For TC intensity, comparisons of FV3\_zc to GFS show that FV3\_zc has largely improved the intensity for intense storms compared to the GFS globally. Since FV3\_zc adopted a nearly identical physics package to that used in the GFS, the improvements should be regarded as coming primarily from the replacement of the dynamical core. Note that the same TC tracker and comparable horizontal resolution were used on all three sets of forecasts. Therefore, it is a very encouraging result that an advanced dynamical core is contributing to the reduction in TC intensity errors in a full physics global model. However, we also found FV3\_zc overpredicts the intensities for strong storms when compared to the observations. This intensity overprediction bias is reduced in FV3\_mp, where the Zhao–Carr scheme is replaced with the GFDL cloud microphysics scheme. Overall, the FV3\_mp shows much smaller globally averaged intensity biases and absolute errors at most lead times compared to both FV3\_zc and the GFS.

For the TC genesis prediction, our results show that FV3\_mp displays the best performance of the three set of forecasts in hit events, false alarms, and missing cases followed by FV3\_zc. This demonstrates that simply using the up-to-date dynamical core but keeping the original physics package the same, the prediction of TC genesis in the model can be improved. Moreover, the upgraded cloud microphysics scheme can further improve the model TC genesis performance with the updated dynamical core.

In addition to evaluating the hit, false alarm, and miss ratios, a novel method was developed to evaluate the performance of model storm genesis based on the lead time lengths of hit events. The comparisons of the maximum OLT (observed genesis lead time) showed that both fvGFS versions predicted TC genesis earlier than GFS in all six basins. Two pairs of comparisons pass the Mann–Whitney  $U$  test: the FV3\_mp shows significantly longer OLT than the GFS in the northeast Pacific basin, and the FV3\_zc shows significantly longer OLT than the GFS in the northwest Pacific basin.

For the model accuracy in generating storms at the observed TC genesis time, our results show a large variation among the basins. None of the three sets of forecasts shows an overwhelmingly higher accuracy in all six basins. However, it was found that the choice of the “observed TC genesis times” can influence the results. If the first record in the ATCF best track data were used as the observed TC genesis times, which could be a disturbance (DB), a low (LO), or a tropical wave (WV), instead of the first tropical depression (TD) record, a substantial number of early events in all three sets of forecasts shift to later time in most basins. It indicates that the model predicted genesis events can sometimes occur during the precursor stage of the observed TC. For TCs in the northeast Pacific and the north Indian basin, the time periods from the first reported “DB” (“LO” or “WV”) to the first reported “TD” were relatively shorter than other basins. Therefore, the ratios of “early” and “late” events did not change too much in the test for these two basins.

The best track data include not only the observed information but also the subjective decisions from experienced forecasters when observed storm structure and intensity estimates are not available. In addition, since the six basins discussed in this study are not covered by the same operational center, different procedures for defining genesis at each center may cause some differences in length of the precursor stage among basins. Therefore, using the first TD records in the best track data as the observed TC genesis times should provide more uniform comparisons between the forecasts rather than using the records during the precursor

stage. However, since the model can predict the occurrence of genesis events during the precursor stage of the observed TC, the comparisons starting from the later records may not fully demonstrate the model's ability to forecast TC genesis.

In regards to the TC genesis intensity forecasts, the two fvGFS versions show larger positive biases than the GFS globally. This overpredicted TC genesis intensity in the fvGFS is consistent with the results of the wind–pressure relationship for all TCs shown in Figs. 5 and 6. The overprediction of TC genesis intensity also caused the relatively large RMSEs in the fvGFS forecasts in most of the basins.

Our results of hit events/false alarms/missing cases and the length of maximum OLT show that fvGFS forecasts outperform the GFS forecast. In addition, the advanced cloud microphysics shows further improvements when investigating the TC genesis forecast. However, it is relatively hard to rank the performance of accuracy in generating storms at the observed TC genesis time among the three set of forecasts. Halperin et al. (2013) discussed forecasts of TC genesis in five global models in the North Atlantic basin. They found that the model's ability to predict TC genesis varies in time and space, and the best-performing model varies from year to year as well. Therefore, only based on cases in a single season, it may be relatively difficult to examine the impacts of a new dynamical core and updated cloud microphysics on the TC genesis forecasts.

As mentioned, predicting changes in storm intensity is a key challenge for current operational global models. This study demonstrates that updating the GFS dynamical core to the nonhydrostatic FV3 can largely improve TC intensity forecasts. Furthermore, we found additional improvements in TC intensity and genesis forecasts when replacing the Zhao–Carr cloud microphysics scheme with the advanced GFDL cloud microphysics scheme. At GFDL we have found that the advection scheme used in the dynamical core has a large impact on TC intensity. The two-delta filter in the non-monotonic advection scheme and the monotonicity constraint in the tracer advection affect the model diffusivity, which can also impact the diabatic heating and the location of the TC deep convection relative to the eye (Harris et al. 2018). Moreover, the individual advection of the six species in the GFDL cloud microphysics scheme compared to the advection of a single condensate species in the Zhao–Carr scheme is a significant difference that can have a major impact on moist processes (Zhou et al. 2019). The above factors also interact with many other processes in a full three-dimensional dynamical model, including the planetary boundary layer scheme, the parameterized convection, and subgrid terrain effects.

The improvements achieved by fvGFS on TC intensity are the fruits of many years of development. The updated dynamical core and advanced cloud microphysics scheme are the two most important factors but may not completely explain the improved results. Based on the results presented in this study, we are confident that with further improvement to the dynamical core and physical parameterizations, the next generation GFS adopted by the NOAA/NWS has great potential to become one of the leading operational TC forecasting systems in the world.

*Acknowledgments.* Special thanks to Tim Marchok for thorough discussions of this work and for the GFDL internal review. We also thank Nathaniel Johnson for the GFDL internal review. We appreciate Dr. Fiorino and two anonymous reviewers for their suggestions. This research was supported under Award NA14OAR4320106 from the National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce.

#### REFERENCES

- Balaji, V., 2012: The flexible modeling system. *Earth System Modelling*, S. Valck, R. Redler, and R. Budich, Eds., Vol. 3, *Springer Briefs in Earth System Sciences*, Springer, 33–41, [https://doi.org/10.1007/978-3-642-23360-9\\_5](https://doi.org/10.1007/978-3-642-23360-9_5).
- Bender, M. A., T. Marchok, C. Sampson, J. Knaff, and M. Morin, 2017: Impact of storm size on prediction of storm track and intensity using the 2016 operational GFDL hurricane model. *Weather Forecasting*, **32**, 1491–1508, <https://doi.org/10.1175/WAF-D-16-0220.1>.
- Chen, J.-H. and S.-J. Lin, 2011: The remarkable predictability of inter-annual variability of Atlantic hurricanes during the past decade. *Geophys. Res. Lett.*, **38**, L11804, <https://doi.org/10.1029/2011GL047629>.
- , and —, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate*, **26**, 380–398, <https://doi.org/10.1175/JCLI-D-12-00061.1>.
- , and —, 2016: Seasonal prediction of tropical cyclones in a new non-hydrostatic GFDL HiRAM. *32nd Conf. on Hurricanes and Tropical Meteorology*, San Juan, Puerto Rico, Amer. Meteor. Soc., 3C.3, <https://ams.confex.com/ams/32Hurr/webprogram/Paper293909.html>.
- , X. Chen, S.-J. Lin, L. Magnusson, M. Bender, L. Zhou, and S. Rees, 2018: Tropical cyclones in GFDL fvGFS—Impacts of DYCORE, physics, and initial conditions. *33rd Conf. on Hurricane and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 9B.4, [https://ams.confex.com/ams/33HURRICANE/webprogram/Manuscript/Paper339827/9B.4\\_extended\\_abstract.pdf](https://ams.confex.com/ams/33HURRICANE/webprogram/Manuscript/Paper339827/9B.4_extended_abstract.pdf).
- , and Coauthors, 2019: Advancements in hurricane prediction with NOAA's next-generation forecast system. *Geophys. Res. Lett.*, **46**, 4495–4501, <https://doi.org/10.1029/2019GL082410>.
- Chun, H.-Y., and J.-J. Baik, 1994: Weakly nonlinear response of a stably stratified atmosphere to diabatic forcing in a uniform flow. *J. Atmos. Sci.*, **51**, 3109–3121, [https://doi.org/10.1175/1520-0469\(1994\)051<3109:WNROAS>2.0.CO;2](https://doi.org/10.1175/1520-0469(1994)051<3109:WNROAS>2.0.CO;2).
- Clough, S. A., M. W. Shephard, E. Mlawer, J. S. Delamere, M. Iacono, K. Cady-Pereira, S. Boukabara, and P. D. Brown,

- 2005: Atmospheric radiative transfer modeling: A summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transfer*, **91**, 233–244, <https://doi.org/10.1016/j.jqsrt.2004.05.058>.
- Delworth, T. L., and F. Zeng, 2012: Multicentennial variability of the Atlantic meridional overturning circulation and its climatic influence in a 4000 year simulation of the GFDL CM2.1 climate model. *Geophys. Res. Lett.*, **39**, L13702, <https://doi.org/10.1029/2012GL052107>.
- , and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643–674, <https://doi.org/10.1175/JCLI3629.1>.
- Donner, L. J., and Coauthors, 2011: The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL Global Coupled Model CM3. *J. Climate*, **24**, 3484–3519, <https://doi.org/10.1175/2011JCLI3955.1>.
- Gao, K., J.-H. Chen, L. Harris, S.-J. Lin, B. Xiang, and M. Zhao, 2017: Impact of intraseasonal oscillations on the tropical cyclone activity over the Gulf of Mexico and Western Caribbean Sea in GFDL HiRAM. *J. Geophys. Res. Atmos.*, **122**, 13 125–13 137, <https://doi.org/10.1002/2017JD027756>.
- Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting*, **28**, 1423–1445, <https://doi.org/10.1175/WAF-D-13-00008.1>.
- Harris, L. M., S.-J. Lin, and C. Tu, 2016: High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *J. Climate*, **29**, 4293–4314, <https://doi.org/10.1175/JCLI-D-15-0389.1>.
- , K. Gao, H. Shin, and A. Hazelton, 2018: Improvements to boundary-layer representation in convection-resolving global-to-regional forecasts using GFDL fvGFS. *33rd Conf. on Hurricane and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 10B.5, <https://ams.confex.com/ams/33HURRICANE/webprogram/Paper340071.html>.
- Hazelton, A. T., M. Bender, M. Morin, L. Harris, and S.-J. Lin, 2018: 2017 Atlantic hurricane forecasts from a high-resolution version of the GFDL fvGFS model: Evaluation of track, intensity, and structure. *Wea. Forecasting*, **33**, 1317–1337, <https://doi.org/10.1175/WAF-D-18-0056.1>.
- Hong, S. Y., and H. L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124**, 2322–2339, [https://doi.org/10.1175/1520-0493\(1996\)124<2322:NBLVDI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<2322:NBLVDI>2.0.CO;2).
- Kim, Y.-J., and A. Arakawa, 1995: Improvement of orographic gravity wave parameterization using a mesoscale gravity wave model. *J. Atmos. Sci.*, **52**, 1875–1902, [https://doi.org/10.1175/1520-0469\(1995\)052<1875:IOGWP>2.0.CO;2](https://doi.org/10.1175/1520-0469(1995)052<1875:IOGWP>2.0.CO;2).
- 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](https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2).
- Marchok, T. P., 2002: How the NCEP tropical cyclone tracker works. *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., P1.13, [https://ams.confex.com/ams/25HURR/techprogram/paper\\_37628.htm](https://ams.confex.com/ams/25HURR/techprogram/paper_37628.htm).
- Miller, R. J., A. J. Scrader, C. R. Sampson, and T. L. Tsui, 1990: The Automated Tropical Cyclone Forecast System (ATCF). *Wea. Forecasting*, **5**, 653–660, [https://doi.org/10.1175/1520-0434\(1990\)005<0653:TATCFS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1990)005<0653:TATCFS>2.0.CO;2).
- Moorthi, S., H. L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. NWS Tech. Procedures Bull. 484, 14 pp.
- Murakami, H., and Coauthors, 2016: Seasonal forecasts of major hurricanes and landfalling tropical cyclones using a high-resolution GFDL coupled climate model. *J. Climate*, **29**, 7977–7989, <https://doi.org/10.1175/JCLI-D-16-0233.1>.
- Pan, H. L., and W.-S. Wu, 1995: Implementing a mass flux convection parameterization package for the NMC medium-range forecast model. NMC Office Note 409, 40 pp.
- Putman, W. 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>.
- Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, [https://doi.org/10.1175/1520-0477\(2000\)081<1231:TATCFS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2).
- Satoh, M., T. Matsuno, H. Tomita, H. Miura, T. Nasuno, and S. Iga, 2008: Nonhydrostatic Icosahedral Atmospheric Model (NICAM) for global cloud resolving simulations. *J. Comput. Phys.*, **227**, 3486–3514, <https://doi.org/10.1016/j.jcp.2007.02.006>.
- Zhao, M., I. M. Held, and G. A. Vecchi, 2010: Retrospective forecast of the hurricane season using a global atmospheric model assuming persistence of SST anomalies. *Mon. Wea. Rev.*, **138**, 3858–3868, <https://doi.org/10.1175/2010MWR3366.1>.
- Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. 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>.