Tropical Cyclone Forecasts in the DIMOSIC Project – Medium-Range Forecast Models with Common Initial Conditions

Jan-Huey Chen[1], Linjiong Zhou[2], Linus Magnusson[3], Ron McTaggart-Cowan[4], and Martin Köhler[5]

[1]NOAA GFDL/ UCAR
[2]Cooperative Institute for Modeling Earth Systems, and Program on Atmospheric and Oceanic Sciences, Princeton University
[3]European Centre for Medium-Range Weather Forecasts
[4]Environment and Climate Change Canada
[5]Deutscher Wetterdienst

December 16, 2022

Abstract

The Tropical cyclone (TC) forecast skill of the eight global medium-range forecast models which are participating in the DIMOSIC (DIfferent Models, Same Initial Conditions) project is investigated in this study. Each model was used to generate 10-day forecasts from the same initial conditions provided by the European Centre for Medium-Range Weather Forecasts. There are a total of 123 initial dates spanning in one year from June 2018 to June 2019 with a 3-day interval. The TC track and intensity forecasts are evaluated against the best track dataset. TC-related precipitation and tropical cyclogenesis forecasts are also compared to explore the differences and similarities of TC forecasts across the models. This comparison of TC forecasts allows model developers in different centers to benchmark their model against other models, with the impact of the initial condition quality removed. The verifications reveal that most models show slow-moving and right-of-track biases in their TC track forecasts. Also, a common dry bias in TC-related precipitation indicates a general deficiency in TC intensity and convection in the models which should be related to insufficient model resolution. These findings provide important references for future model developments.

Tropical Cyclone Forecasts in the DIMOSIC Project -- Medium-Range Forecast Models with Common Initial Conditions

Jan-Huey Chen[1,2], Linjiong Zhou[3], Linus Magnusson[4], Ron McTaggart-Cowan[5], and Martin Köhler[6]

[1]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

[2]University Corporation for Atmospheric Research, Boulder, CO, USA

[3]Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, NJ, USA

[4]European Centre for Medium-Range Weather Forecasts, Reading, UK

[5]Environment and Climate Change Canada, Montreal, Canada

[6]Deutsche Wetterdienst, Offenbach, Germany

Corresponding author: Jan-Huey Chen (Jan-Huey.Chen@noaa.gov)

Key Points:

- Tropical cyclone forecasts are compared between global medium-range models from leading modeling centers initialized with identical data.
- Similarities and differences between the models set a benchmark of TC forecast with the impact of the initial condition quality removed.
- Common TC forecast biases indicate general deficiencies in the models and suggest a direction for further model improvement.

Abstract

The Tropical cyclone (TC) forecast skill of the eight global medium-range forecast models which are participating in the DIMOSIC (DIfferent Models, Same Initial Conditions) project is investigated in this study. Each model was used to generate 10-day forecasts from the same initial conditions provided by the European Centre for Medium-Range Weather Forecasts. There are a total of 123 initial dates spanning in one year from June 2018 to June 2019 with a 3-day interval. The TC track and intensity forecasts are evaluated against the best track dataset. TC-related precipitation and tropical cyclogenesis forecasts are also compared to explore the differences and similarities of TC forecasts across the models. This comparison of TC forecasts allows model developers in different centers to benchmark their model against other models, with the impact of the initial condition quality removed. The verifications reveal that most models show slow-moving and right-of-track biases in their TC track forecasts. Also, a common dry bias in TC-related precipitation indicates a general deficiency in TC intensity and convection in the models which should be related to insufficient model resolution. These findings provide important references for future model developments.

Plain Language Summary

Despite recent improvements in our ability to predict the track and intensity of tropical cyclones, these storms remain significant forecasting challenges. Forecasters rely heavily on the guidance generated by numerical weather prediction systems, making the reliability of these systems essential for accurate forecasts during these high-impact weather events. As a result, improvement the quality of tropical cyclone guidance is an important numerical model development objective. In this study, the TC forecast skills in the eight global medium-range forecast models from the model development centers/institutes who participated in the DIMOSIC (DIfferent Models, Same Initial Conditions) project are examined. All models were initialized from the same data provided by the ECMWF (European Centre for Medium-Range Weather Forecasts) to investigate the differences and similarities among their TC forecasts without the impact of the quality of initial conditions. Besides the general TC forecast evaluation metrics including errors and biases of the track and intensity, the TC-related precipitation and TC genesis skills are also evaluated to comprehensively explore the performance of TC forecasts among all models. The comparison allows model developers in different centers to benchmark their model against other participating models. Moreover, the verification results provide important references for future model developments.

1 Introduction

Tropical cyclone (TC) prediction is an important mission for weather and climate agencies in many countries. Over the past few decades, numerical models have become the most important tools for operational centers to make TC forecasts on weather and sub-seasonal to seasonal time scales. Therefore, improving the model performance of TC forecasts has been one of the leading tasks in most operational centers or modeling research institutes working on model development. In addition, the accurate depiction of physical processes that lead to a better TC forecast in the model are also relevant to interesting scientific questions in the atmospheric science research area more broadly.

The quality of initial conditions has a leading impact on short- to medium-range forecast skill, including for TC forecasts. In Chen et al. (2019a), the fvGFS (finite volume Global Forecasting System) model developed at the Geophysical Fluid Dynamics Laboratory (GFDL) initialized with the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) data showed much-improved TC track forecasts for the 2017 Atlantic hurricane season compared to its retrospective forecasts initialized with the data from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) version 14. In Magnusson et al. (2019), the same approach was used, comparing the GFDL fvGFS model forecasts to those from the IFS and GFS. The results showed that the choice of initial conditions clearly dominated the forecast quality in the medium-range predictions, but that the model formulation could also play a significant role.

Since major model development centers mostly develop their modeling systems independently, the DIMOSIC project (DIfferent Models, Same Initial Conditions; Magnusson et al. 2022) was devised to investigate the relationship between the choice of model formulation and forecast quality. Models developed by different world-leading modeling centers were initialized from the same initial condition. In Magnusson et al. (2022), the differences and similarities of the forecasts among the models were presented. The results found that some pairs of models behaved more similarly than other pairs due to their sharing of partial physical parameterizations, e.g. ECWMF IFS and DWD (Deutsche Wetterdienst) ICON (Icosahedral Non-hydrostatic Model). On the other hand, ICON and GFDL SHiELD (System for High-Resolution Prediction on Earth-to-Local Domains) showed relatively large forecast differences, while both ranking among the best models. Regarding the influences from model formulations on the forecasts, however, it was difficult to point out a single model component that had the strongest impact on the forecast differences. Also, as pointed out by Magnusson et al. (2022) the interaction between different model parameterizations and their respective configurations could play a significant role as well.

In this study, the performance of TC forecasts from the DIMOSIC models is evaluated. The TC track and intensity forecast skills among the models during the period of June 2018 to June 2019 are compared. Since TC intensity in interpolated data does not reflect the actual TC intensity at the native model resolution, the TC-related precipitation are also evaluated to provide another perspective on forecasted TC activities for better exploring the differences and similarities among the models. Also, the forecast skill of TC genesis was investigated by comparing the hit/false alarm ratios among the models, as well as using the method based on the lengths of TC genesis lead time introduced in Chen et al. (2019b) to examine the accuracy of TC genesis timing in the model forecasts. These comparisons should be valuable for model developers in different centers to benchmark their model's performance on TC forecasts against that of other models, with the impact of the initial condition quality removed.

The models participating in the DIMOSIC project are introduced in section 2 which also describes the observation data and methodology used in this study. The comparisons of track, intensity, TC-related precipitation, and genesis forecasts among the models are contained in section 3. Summary and discussion are presented in section 4.

2 DIMOSIC models, forecasts, and verification data

General information on the numerical models and their developing centers/institutes participating in the DIMOSIC project are listed in Table 1. The horizontal resolutions and the number of vertical levels of the models and their key references are included. Some centers/institutes submitted more than one model configurations to the project, but we only investigate one configuration of the model for each center/institute based on their suggestions. The only exception is to include two versions of IFS 45R1 and 47R3, to provide an example of the incremental change obtained for an upgrade of one model. For the sea surface temperature evolution in the models, the two IFSs used a partial coupling to the 3D ocean NEMO model (Mogensen et al. 2017), SHiELD is coupled with a 1D mixed layer ocean model (Pollard et al. 1973), CMC used a thermodynamic mixed layer ocean model (Zeng and Beljaars 2005), and others used persistent anomalies from the analysis. Other detailed configurations of each model including dynamical cores and major physical parameterizations can be found in the sub-section of “Model descriptions” in the section of “Models and data“ and in Table 2 in Magnusson et al. (2022), but is not repeated in this paper.

Acronyms	Models	Centers/Institutes	Resolution	Key references
ARPEGE	Action de Recherche Petite Echelle Grande Echelle (version: 46T1)	Meteofrance	5-25 km 105 levels	Roehrig et al. (2020)
CMC	Global Environmental Multiscale Model (GEM) (version: v5.0.2)	Canadian Meteorological Center (CMC)	15 km 80 levels	Girard et al. (2014) McTaggart-Cowan et al. (2019)
ICON	Icosahedral Non-hydrostatic Model (version: April 21)	Deutsche Wetterdienst (DWD)	13km 90 levels	DWD (2022)
IFS/ IFS-47R3	Integrated Forecasting System (versions: 45R1 and 47R3)	European Centre for Medium-range Weather Forecasts (ECMWF)	9 km 137 levels	ECMWF (2018, 2021)
JMA	Global Spectral Model (GSM) (version: GSM1705)	Japan Meteorological Agency (JMA)	20 km 100 levels	JMA (2019)
SHiELD	System for High-Resolution Prediction on Earth-to-Local Domains (version: rt2019)	Geophysical Fluid Dynamics Laboratory (GFDL)	13 km 91 levels	Harris et al. (2020)
UM	Unified Model	UK Met Office	10 km 70 levels	Walters et al. (2019)

TABLE 1. The eight DIMOSIC models (the ECWMF contributed two IFS configurations). From left to right: model acronyms, model full names (and versions if applicable), centers/institutes that models belong to, horizontal resolutions and the number of vertical levels used in the models, and the key references of the models.

All models conducted 10-day forecasts from the same initial conditions: ECMWF operational analyses based on the IFS model cycle 45R1 (ECMWF 2018). The 9-km analyses on 137 vertical levels are generated from a 4DVar data assimilation system (Rabier et al. 2000). All participating institutes received the interpolated data at 0.1 degree for their model initialization. Detailed procedures for handling the initial conditions in each model are described in section 2c and Table 3 of Magnusson et al. (2022).

The total 123 forecasts are conducted with initialization dates spanning one year from June 2018 to June 2019 at 3-day intervals. The 10-day forecast outputs from each model were interpolated to a common 0.5 degree grid using an average interpolation method available in the EcCodes/MIR package (https://confluence.ecmwf.int/display/ECC/ecCodes+Home). The GFDL simpler tracker (Harris et al. 2016) was used with a warm-core criterion to track TCs in the forecasts of the eight models based on the fields of sea-level pressure, 10-m wind speed, 850-hPa vorticity, and mean temperature between 500-300 hPa.

There were 109 observed TCs in the DIMOSIC period. Their storm tracks based on the best track data (b-deck) in Automated Tropical Cyclone Forecast (ATCF) dataset (Miller et al. 1990; Sampson and Schrader 2000) are shown on the map in Fig 1. There were 35 TCs in the northwest Pacific basin (WPAC), 24 in the northeast Pacific basin (EPAC), and 16 in the North Atlantic basin (NATL). In the Southern Hemisphere (SHEM), there were 27 TCs in the combined South Indian Ocean and South Pacific Ocean. Besides the overview of global analyses, the individual TC forecast skill in the above four major sub-regions will be investigated and compared with each other in the following sections.

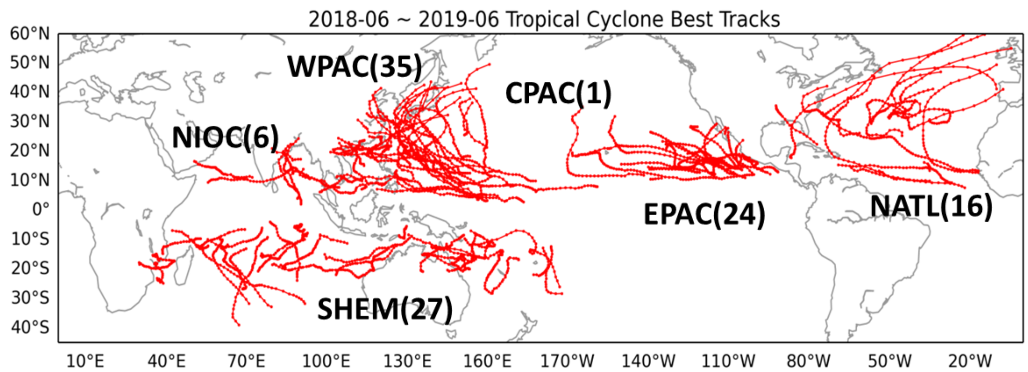

FIGURE 1. All TCs in the DIMOSIC period. The best tracks are from the ATCF dataset. The numbers of TCs are indicated in the brackets next to the acronyms of the six regions: WPAC: northwest Pacific basin; EPAC: northeast Pacific basin; CPAC: north-central Pacific basin; NATL: North Atlantic basin; NIOC: north Indian Ocean; SHEM: Southern Hemisphere.

The ATCF dataset was used to evaluate the forecast errors of TC track and intensity, and the skill of TC genesis forecasts in the models at six-hour intervals. For the TC-related precipitation, the NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) dataset was used (Hong et al. 2004) for verification. To equally compare to the model output field of total precipitation, this high-resolution (0.1 degree) satellite observational dataset was interpolated into 0.5 degree.

3. Results

3.1 TC track forecasts

The prediction of TC path, or track, is the one of the most important factors for taking necessary precautions against possible impacts from hurricanes or typhoons. The homogeneous comparisons of global mean TC track forecast errors along with the forecast lead time at 12-h intervals are shown in Fig. 2a. The differences among the models are small through the 72-hour lead time with the exception of CMC which shows a relatively higher error than others. During the 72 to 120-hour head time, TC track errors diverge into two groups. The two IFSs, ICON, SHiELD, and ARPEGE show lower errors than UM, CMC, and JMA. Both versions of IFS show the lowest TC track errors after the 120-hour lead time all the way to the 168-hour lead time, followed by ICON and SHiELD. Note that most models are within the 95% confidence levels of IFS, which means the differences in forecast skills are not statistically significant. To highlight the difference between the leading models, Fig. 2b shows the differences in TC track errors of five of the models compared to the IFS. The newer IFS-47R43 performs slightly better (negative values in track error differences) or equivalently during the entire 7 days. In the early lead times (36-84 hours), most of the five models also show slightly better forecasts than IFS. ICON displays similar (or slightly better) skill to the two IFS versions until the 120-hour lead time. Both SHiELD and ARPEGE perform very well before the 96-hour lead time. After the 120-hour lead time, SHiELD shows lower forecast errors than other models except for the two IFSs.

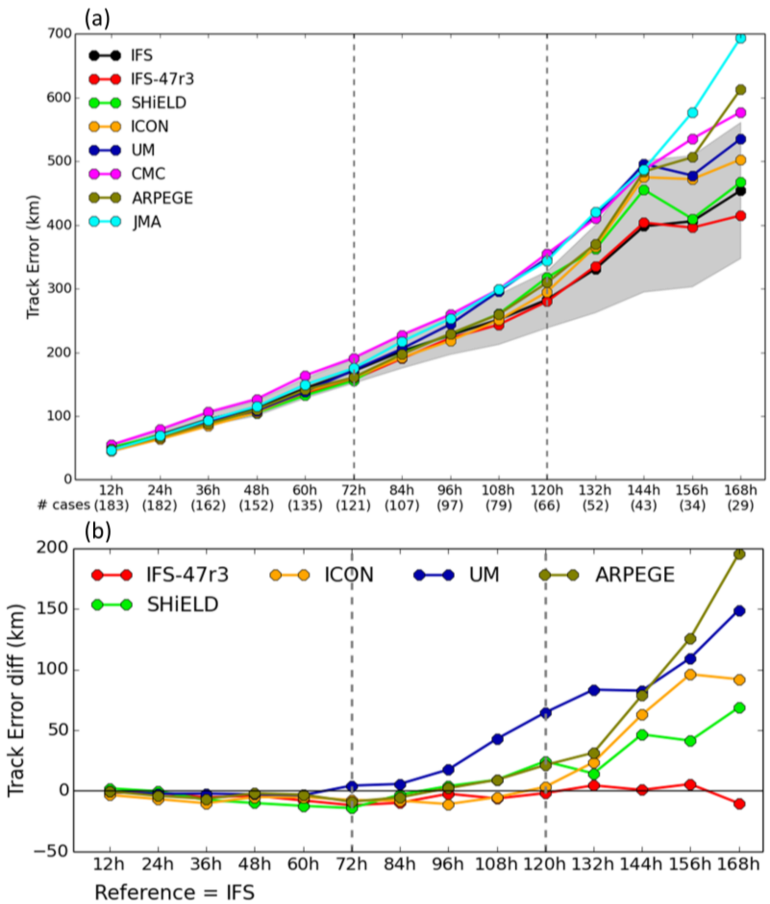

FIGURE 2. (a) Global mean TC track forecast errors (km) at every 12 hour forecast lead time for IFS (black), IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), CMC (magenta), ARPEGE (grass green), and JMA (light blue). The 95% confidence levels for IFS are indicated by the gray color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom of each abscissa. Vertical gray dotted lines indicate 72 and 120 hour forecast lead times. (b) Global mean TC track forecast error differences of IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), and ARPEGE (grass green) comparing to IFS.

Figure 3 shows the 5-day average TC track errors for all models in the entire globe and in the four major sub-regions individually. For the two IFSs, IFS-47R3 shows lower track errors than IFS globally and in all major sub-regions. ICON shows competitive low track errors to IFS-47R3 in the WPAC and the SHEM, and a very low track error in the NATL. SHiELD performs the lowest track error in the EPAC, and low track errors besides the two IFSs and ICON in the SHEM and the WPAC. However, SHiELD has a much larger track error in the NATL, which results from the slow moving bias shown in the forecasts of Hurricane Florence and the bias of direction of motion shown in the forecasts of Hurricane Leslie. Detailed investigations can be found in Text S1 and Figs. S1-S3. The performance of the TC track forecast in UM is notable.

The model shows the lowest track error in the NATL but the largest track error in the EPAC, while its track errors in the WPAC and the SHEM are also at the high end compared to other models. JMA performs competitively to IFS-47R3 and ICON in EPAC, but not in other sub-regions. With regard to CMC, it shows larger track errors than other models in most sub-regions except for in the EPAC during this targeted year. From Fig. 3, we can also see that the globally-averaged track forecast errors among the models is dominated by the errors in the WPAC and the SHEM since the majority of the TCs were found in these two sub-regions.

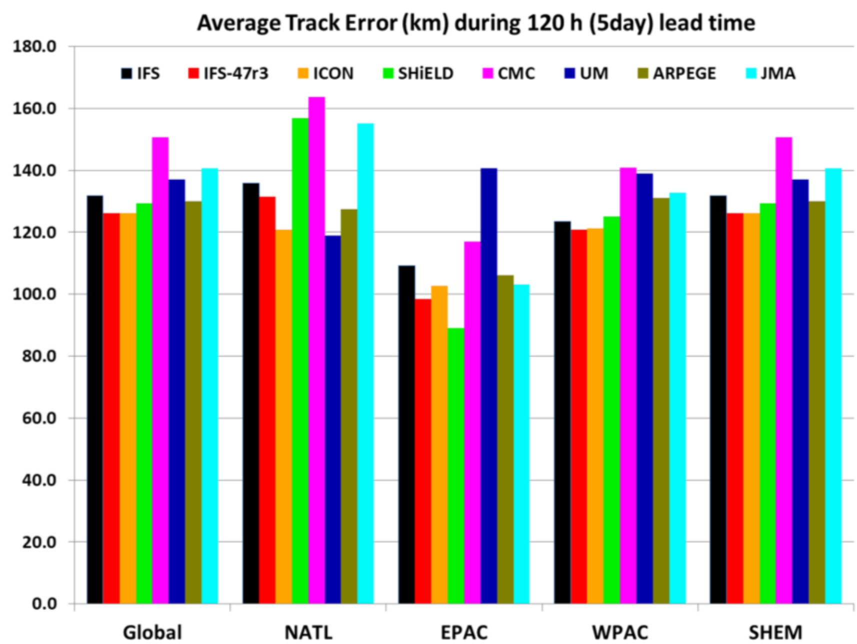

FIGURE 3 Averaged Track errors (km) in globe and 4 sub-regions during the 120-hour lead time for the 8 models. Abbreviations and colors used for the models are the same as in Fig. 2a. Abbreviations used for the sub-regions on the abscissa are the same as in Fig. 1.

The sources of track errors can be due to biases either in forecasts of the TC translational speed or the TC direction of motion. The lower sub-panels in Fig. 4 show the globally averaged along-track (AT) errors and cross-track (CT) errors (perpendicular to the track) for all models. Both AT and CT errors are calculated as great circle distances. Most of the models start to show negative AT biases and positive CT biases during 72 to 120-hour lead time, which indicates that the TC track errors during the later lead times are mostly due to the slow and northward (for an easterly moving TC) moving biases. In general, UM shows the smallest AT and CT biases among all models, and the biases of SHiELD take place at longer forecast lead times than other models.

By the Pythagorean Theorem the square of the total error equals the squares of the AT and CT errors (Chen et al. 2019a). The squares of total track errors, CT errors, and AT errors are plotted in upper sub-panels in Fig. 4 to illustrate the proportion of contributions from the AT and CT errors respectively to the total error. From Figs. 4a,b, we can see that the AT error contributes more than the CT error to the total track error in the two IFSs, which indicates that the track errors in the IFSs are dominated by the slow-moving bias (negative AT biases). The characteristic of the consistently larger AT error square than the CT error square can also be found in ICON, CMC, and UM, but the differences between their AT and CT error squares are smaller than those in the IFSs. SHiELD, ARPEGE, and JMA show relatively closer AT and CT error squares, especially SHiELD. This indicates that the contributions of the slow and northward moving biases to the total track errors are similar in these models. For ARPEGE, the CT error is the dominant track error in late lead times, while the AT error contributes more to the JMA's total TC track errors during the 120 to 144-hour lead time.

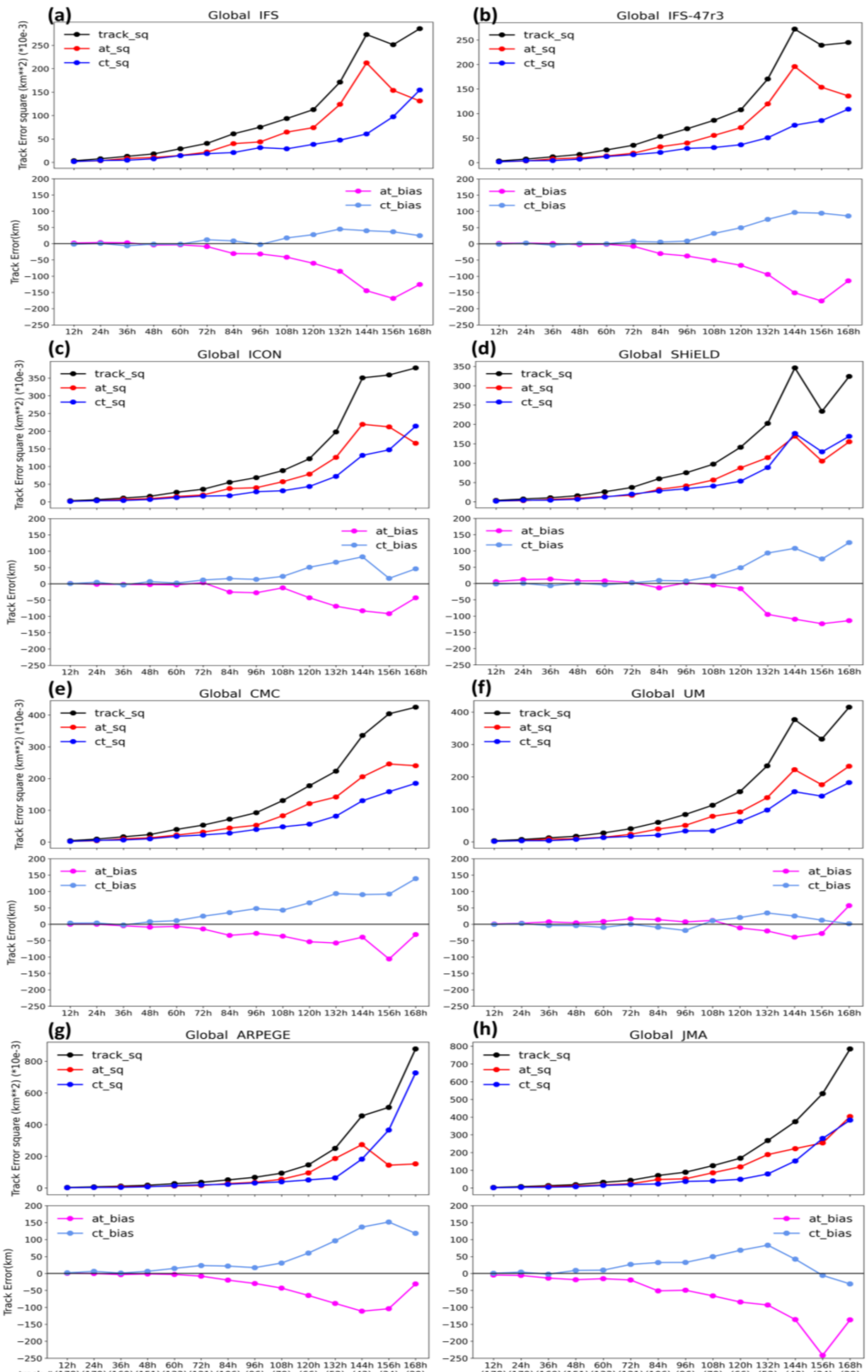

FIGURE 4 Global analyses of along-track (AT) error and cross-track (CT) error for (a) IFS, (b) IFS-47R3, (c) ICON, (d) SHiELD, (e) CMC, (f) UM, (g) ARPEGE, and (h) JMA. The squares of total track errors (black), along-track errors (red), and cross-track errors (blue) are in the upper panels for each model. The biases of along-track (magenta) and cross-track (light blue) errors are in the lower panels. Numbers of homogeneous cases for each lead time are listed at the bottom of lower panels.

It is also found that models have different AT and CT errors in different sub-regions. All models showed a slow-moving bias and a poleward bias in the NATL and the WPAC, but in the EPAC, except for JMA, most models show a fast-moving bias. In contrast, there are no consistently slow or fast moving biases among models in the SHEM. Detailed analyses of AT and CT errors for all eight models in the four major sub-regions can be found in Text S2 and Figs. S4-S7.

3.2 TC intensity forecasts

It has been more challenging to predict TC intensity than track, especially for global models which usually cannot resolve fine scale interactions between thermal dynamics and dynamics due to insufficient resolutions. As outlined in Section 2, the 10-day forecast outputs were interpolated to a common 0.5 degree for each model. Therefore, the model-predicted TC intensities found by the tracker are underestimated due to the low data resolution. However, it is still of interest to compare the relative differences of TC intensities among the models for the interpolated output data. The global mean TC intensity errors and biases based on the maximum 10-m wind speed are presented in Fig. 5a. SHiELD predicts a much stronger TC intensity than other models, followed by UM and ARPEGE. Figure 5b uses SHiELD as an example to demonstrate the differences between the TC intensities obtained from the native resolution (13 km grid) outputs and in the interpolated 0.5 degree resolution data. The average differences of total error and bias are between 3 to 5 ms^{-1}.

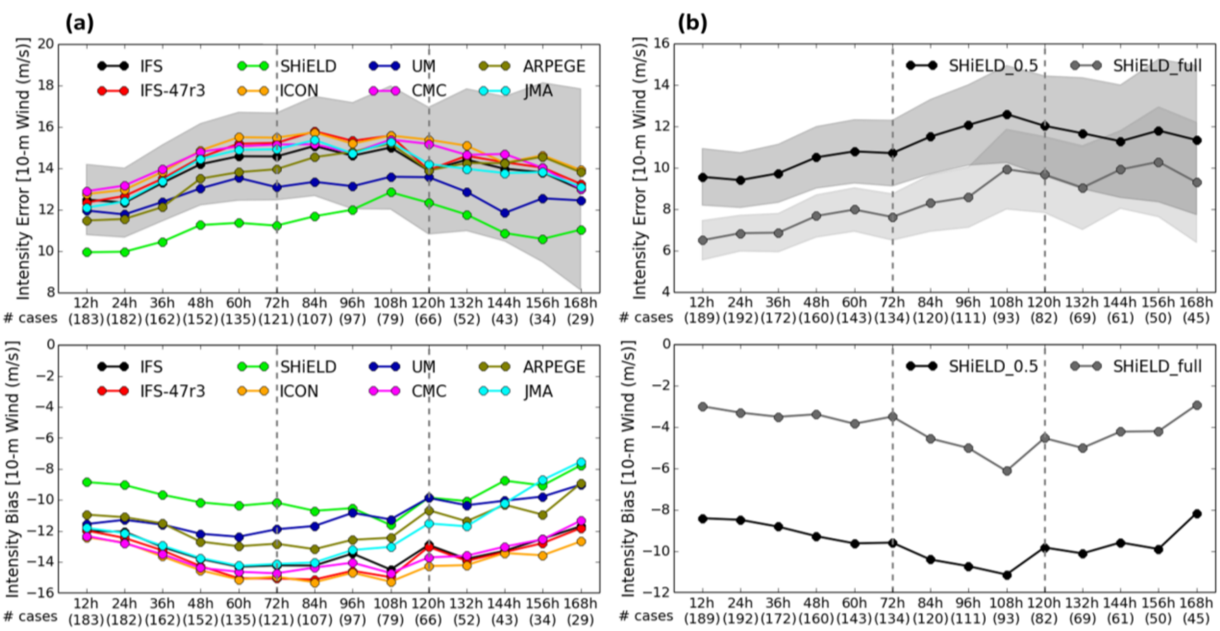

FIGURE 5. (a) Global mean TC intensity errors and biases. Upper panel: Absolute error of the maximum 10-m wind speed ($m\ s^{-1}$) along with the model forecast lead time for 8 models. Abbreviations and colors used for the models are the same as in Fig. 2a. The 95% confidence levels for IFS are indicated by the gray color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom. Vertical grey dotted lines indicate 72 hour and 120 hour lead times. Lower panel: As in the upper panel, but for the bias of the maximum 10-m wind speed ($m\ s^{-1}$). (b) As in (a), but for SHiELD native resolution data (black; SHiELD_full) and SHiELD 0.5 degree interpolated data (gray; SHiELD_0.5). The 95% confidence levels for each resolution data are indicated by the same medium and light transparent grey shading areas, with their overlapping region denoted by dark grey shading.

3.3 Forecasts of TC-related precipitation

Since the performance of TC intensity forecasts cannot be fully represented by the interpolated data, here, the TC-related precipitation is evaluated to provide another perspective on the forecasted TC characteristics in the models. Using the TC track information, the precipitation within 350 km of each TC center is used to investigate the TC-related precipitation for each model. Figure 6 shows the accumulated total precipitation for all TCs during the DIMOSIC period in each model compared to the Global Precipitation Measurement (GPM) observational data (Fig. 6a). The comparison shows that all models under-predict the amount of precipitation, especially in the most active areas of the WPAC and the EPAC. From a broad visual comparison, UM and SHiELD appear to have produced the highest and lowest amounts of precipitation among all the models, respectively. This can be confirmed by comparing the accumulated precipitation of models to the GPM data presented in Fig. 7a. The UM shows a much larger amount of precipitation than other models, followed by ARPEGE and CMC. In contrast, SHiELD shows the least TC-related precipitation among all models, except in the EPAC. The ranks of models are similar in different sub-regions, while the global precipitation amounts are dominated by those in the WPAC and the SHEM, as expected.

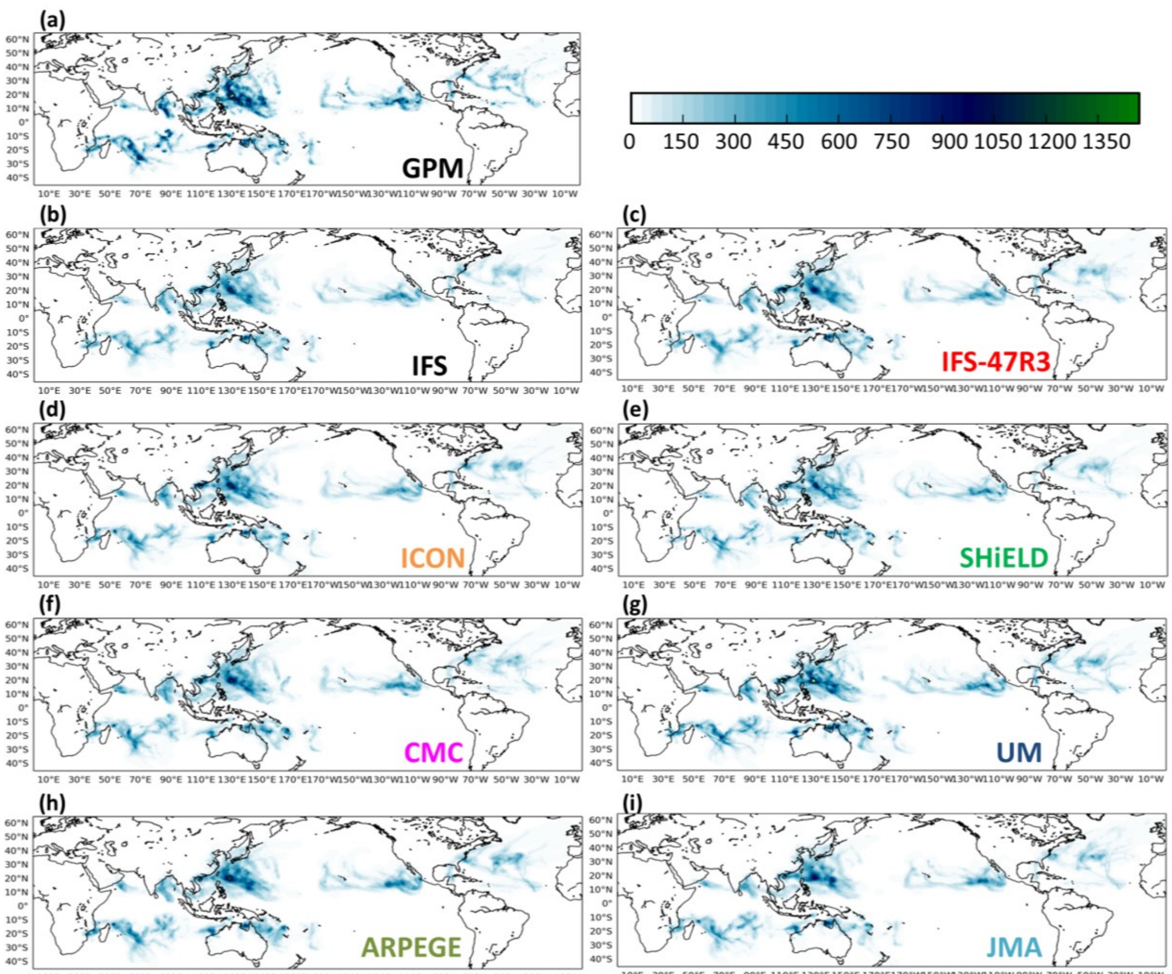

FIGURE 6. Accumulated TC-related precipitation (unit: mm) for all TCs in the DIMOSIC period in (a) Global Precipitation Measurement (GPM) observations, (b) IFS, (c) IFS-47R3, (d) ICON, (e) SHiELD, (f) CMC, (g) UM, (h) ARPEGE, and (i) JMA.

To more objectively compare the forecasted locations of TC-related precipitation in each model to the GPM observations, the equitable threat scores (ETSs; Schaefer 1990) are computed. The ETSs for all TC-related precipitation areas in the four sub-regions for all eight models are compared in Fig. 7b. Note that although UM shows the closest precipitation amount to the GPM observation data (Fig. 7a), its ETS (skill) is lower than other models when considering all TC-related precipitation areas. This could be related to its relatively larger track errors (Fig.3) that cause the displacement of precipitation locations. However, for the areas with at least 300 mm of accumulated TC-related precipitation, the ETSs of UM are generally higher than those of other models (Fig. 7c). This is likely due to its relatively better prediction of precipitation amounts (Fig. 7a). In contrast, SHiELD under-predicts the precipitation amounts, but owing to its better track forecast in the EPAC (Fig. 3), it is able to achieve relatively higher ETSs in this sub-region (Figs. 7b,c). When comparing the two IFSs, Fig. 7 shows that their accumulated precipitation amounts are similar, but the newer IFS-47R3 generally had higher ETS scores (Figs. 7b,c). Finally, JMA shows relatively higher ETSs in the WPAC in both categories, while in the NATL, the highest ETSs in the two categories are achieved by ICON and ARPEGE (Figs. 7b,c).

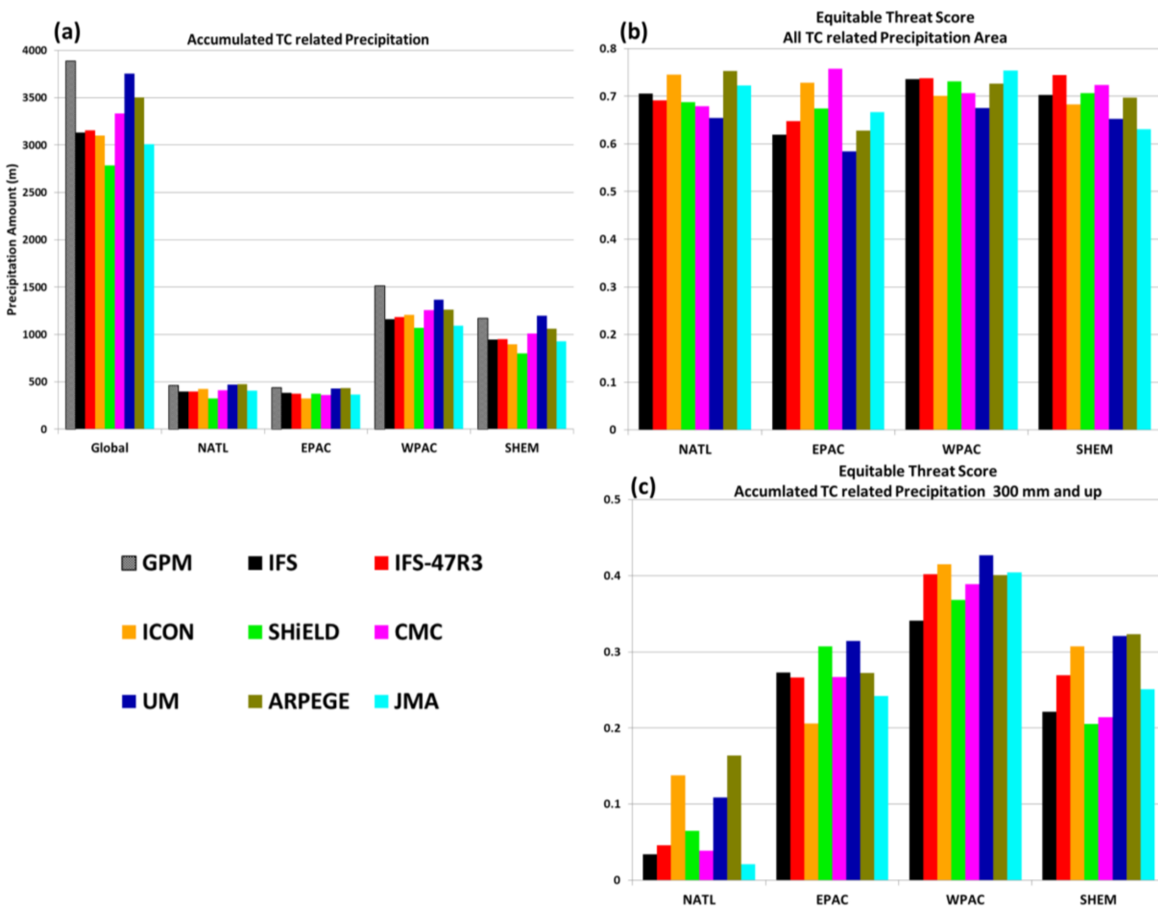

FIGURE 7. (a) Global accumulated TC-related precipitation (unit: m) and for the four sub-regions for all TCs in the DIMOSIC period. The equitable threat scores (ETSs) for (b) all TC-related precipitation areas and (c) the areas with 300 mm and up accumulated TC-related precipitation. The GPM analysis data in (a) is shown in the bars with the grey checkerboard pattern. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 3.

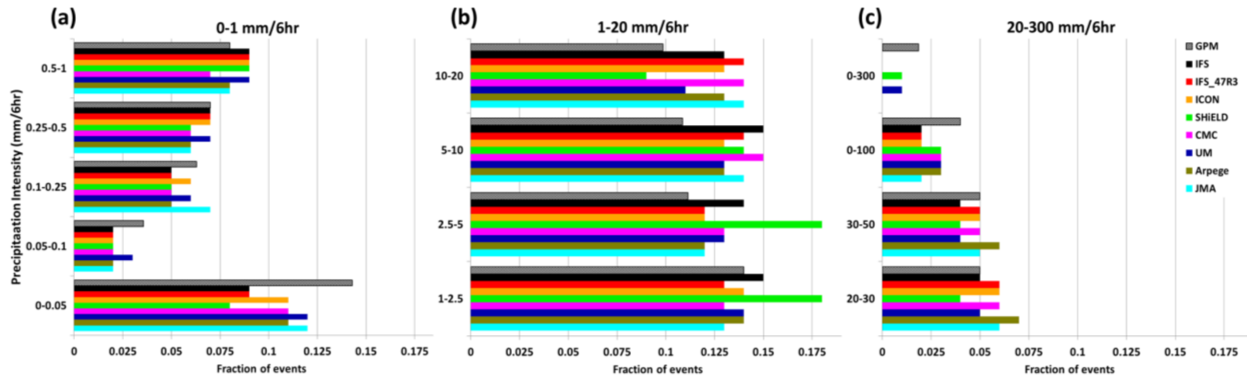

FIGURE 8. Fractions of precipitation events (GPM observations and model forecasts) in each precipitation intensity bin. (a) 5 bins for precipitation intensities from 0 to 1 mm(6 h)$^{-1}$. (b) four bins for precipitation intensities from 1 to 20 mm(6 h)$^{-1}$. (c) four bins for precipitation intensities from 20 to 300 mm(6 h)$^{-1}$. Abbreviations and colors used for the model are the same as in Fig. 7.

TC-related precipitation based on different precipitation intensities was also analyzed. Figure 8 shows the fractions of precipitation events in different precipitation intensity bins. Most of the models under-predicted light (weaker than 0.5 mm (6 h)$^{-1}$; Fig. 8a) and heavy (stronger than 50 mm (6 h)$^{-1}$; Fig. 8c) precipitations, but over-predicted medium precipitation events (Fig. 8b). Although SHiELD and UM were able to predicts some heavy precipitation events in the bin of 100-300 mm (6 h)$^{-1}$, SHiELD significantly over-predicted the events between 1-5 mm (6 h)$^{-1}$. The SHiELD development team at GFDL will closely examine the precipitation forecasts in the model in the near future, particularly to better isolate the possible reasons for these excessive precipitation amounts.

3.4 Forecasts of TC genesis

When a timeline contains an observed TC genesis, the track and intensity forecasts of the TC are verified based on the model forecasts initialized at or after the observed TC genesis time. In contrast, to investigate the models' performance for TC genesis, the 10-day forecast runs initialized before the observed TC genesis time which is based on the first "TD (tropical depression)" recorded in the ATCF best track data are considered. All TCs found by the GFDL simple tracker in these forecasts but not existing as TCs in the initial conditions for the forecast are counted as genesis events in the models. If a TC genesis has a track that "matches" an observed TC track, the genesis case is categorized as a "hit event". Otherwise, it is a "false alarm". The same criteria is used as in Chen et al. (2019b) to judge a model storm was a successful prediction of an observed genesis event.

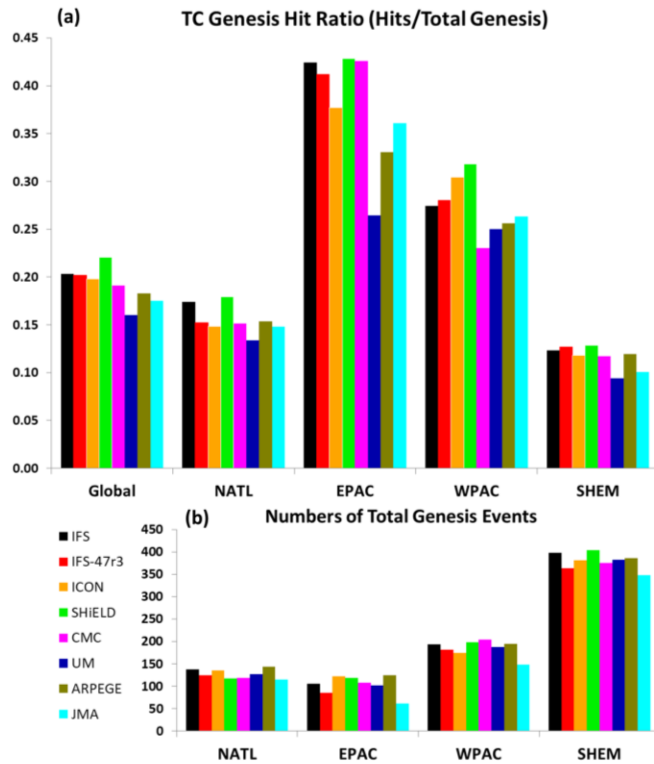

FIGURE 9. (a) Ratios of hit events to the total number of genesis events and (b) Numbers of total genesis evens (sum of hit events and false alarms) for all models. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7.

Figure 9 shows the TC genesis ratios (hits to total predicted genesis events) and the number of total genesis events for each of the eight models in different regions. The sum of hit events and false alarms is equal to the number of total forecasted genesis events. We first find that all models show the highest hit ratios (Fig. 9a) with the fewest total genesis events in the EPAC. This indicates that models can predict TC genesis more skillfully in the EPAC than in other sub-regions. In contrast, models show the lowest hit ratios but the largest numbers of total genesis events in the SHEM, which indicates that models generate more false alarms in this sub-region than in the others. In general, SHiELD demonstrates the highest hit ratios both globally and in all sub-regions, followed by the two IFSs and ICON. CMC also shows high hit ratios in the EPAC. UM shows the lowest hit ratios in most regions with the exception of the WPAC.

During the DIMOSIC period, there were 16, 23, 34, and 27 TCs generated in the NATL, the EPAC, the WPAC, and the SHEM, respectively. However, not all observed TC geneses were predicted by the models. Figure 10a lists the number of TCs which were completely missed by models in each sub-region. It shows that JMA missed the genesis of total 30 TCs globally which is the most among all models. Most TC missed by JMA were in the WPAC and the SHEM. UM also missed many TC geneses in the EPAC. SHiELD had the least number of missed TCs globally followed by IFS and CMC. The newer IFS-47R3 missed more TCs than IFS in the NATL and the EPAC.

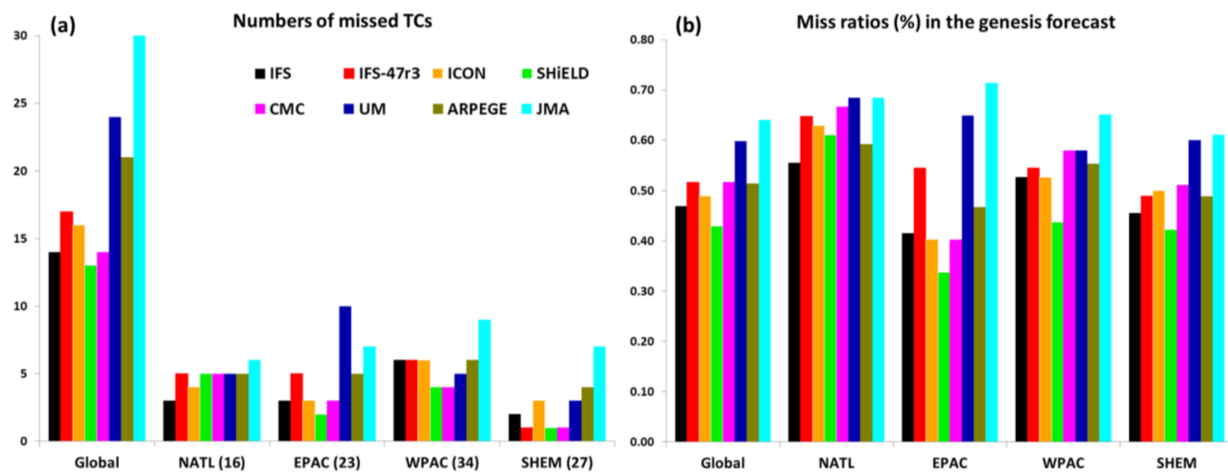

FIGURE 10. (a) Numbers of missed TCs and (b) miss ratios (%) in the genesis forecasts from the eight models. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7. The numbers on the abscissa in (a) indicate the observed TC numbers in each sub-regions.

In the DIMOSIC period, models were initialized every three days. Hence, during the 10 days before an observed TC genesis event, a model could have three or four 10-day forecasts initialized and these runs are expected to predict this genesis event. Therefore, besides counting the number of completely missed TCs, the “miss ratio” can be computed as the number of missing cases compared to the number of expected genesis hit events (Chen et al. 2019b). The miss ratios for each model in the different sub-regions are shown in Fig. 10b to better revel the differences among the models. It shows that SHiELD shows the lowest miss ratios generally, except for in the NATL, while JMA and UM still struggle with relatively high miss ratios globally. We note that IFS-47R3 shows higher miss ratios than IFS in all sub-regions (Fig. 10b), including in the WPAC and the SHEM where IFS-47R3 shows the same or fewer numbers of completely missed TCs than IFS (Fig. 10a).

Following Chen et al. (2019b), beyond the scores of hit events, false alarms, and missing cases, we also investigate how precisely a model could predict the timing of TC genesis by comparing the “length of lead time” (see Fig. 8 in Chen et al. 2019b). The observed genesis lead time (OLT) is defined as the difference in time between the model initial time and the time at which observed TC genesis occurred. On the other hand, the time span from the model initial time to the model-predicted TC genesis lead time is referred as the model genesis lead time (MLT). The differences between the MLT and OLT (DMO) can indicate how accurate a model is in generating storms at the observed genesis time. If a model-predicted TC genesis occurred exactly at the observed TC genesis time, the DMO of this hit event is “zero”. A positive DMO means that the model hit event occurs later than the observed TC genesis time, while negative DMO values are associated with early initiation of the TC in the model

For each observed TC, it is expected that more than one hit event will happen in the set of 10-day forecasts that cover the observed genesis time. To assess the predictive skill of each model, we only consider the maximum OLT, corresponding to the integration that identified the observed TC at the longest lead time. Figure 11a shows the mean values of the maximum OLT of all observed TCs in the four major sub-regions. In the NATL, both SHiELD and ARPEGE show a 110-hour OLT which is longer than the OLTs of other models, e.g. 78-hour OLT of UM. This indicates that SHiELD and ARPEGE could, on average, predict a hit TC genesis event 32 hours earlier than UM in the NATL. SHiELD also shows the earliest hit events in the EPAC and the WPAC, while ARPEGE shows the earliest hit events in the SHEM. The models in this study generally predict hit events earlier in the EPAC than in other sub-regions, except for JMA, which performs betters in the WPAC and the SHEM than in other sub-regions. It is also interesting to see that IFS shows earlier hit events than the newer IFS-47R3 in most sub-regions.

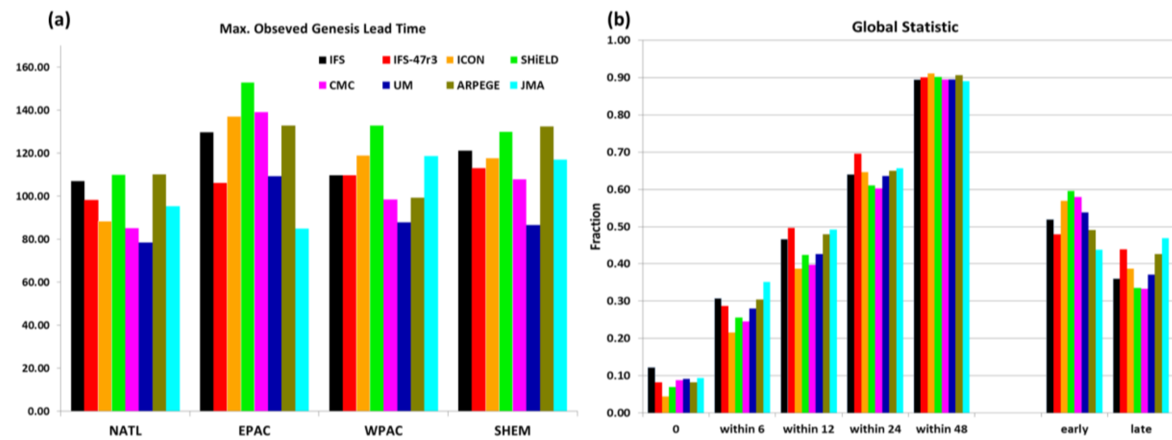

FIGURE 11. (a) Mean values of maximum observed genesis lead time (in hours) of all storms in each sub-region for the eight models. Abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7. (b) Fractions of global total hit events in each model that occurred within a certain DMO length. On the abscissa, “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” is for all hit events with negative DMOs and “late” is for all hit evens with positive DMOs. Abbreviations and colors used for the models in both (a) and (b) are the same as in Fig. 7.

Figure 11b shows the fraction of global total hit events in each model which occurred within a certain length of DMO (indicated on the abscissa). From the definition of DMO, a model showing more genesis cases with short DMOs indicates that the model could predict more accurate genesis timings of its hit events. It can be found that IFS shows the highest fraction among all of the eight models in the “zero” DMO length category. The results indicate that IFS shows the highest ratio of its hit events forecasted at the observed TC genesis time among the models. Besides the two IFSs, JMA and ARPEGE also accurately predict TC genesis timings within the first three categories (“zero”, “within 6”, and “within 12”) which is the 24-hour window (12 hours before or after) centered on the observed TC genesis time. In contrast, ICON shows the smallest fractions in the first three DMO length categories, which implies that the accuracy of TC genesis timing of ICON is relatively low compared to the other models.

From the result of the “within 48” DMO in Fig. 11b, we can see that more than 89% of hit events in each of the models occur within the 48 hours before or after the observed TC genesis time. When comparing the results in the “early” and “late” categories, it is seen that most models forecast their hit events before the observed TC genesis time, except for JMA. The ratios of “early” to “late” cases are larger in SHiELD and CMC compared to other models, while the IFS-47R3 shows relatively even number of cases of hit events generated before or after the observed TC genesis time.

4 Summary and Discussion

The DIMOSIC project provides a great opportunity to engage the worldwide community of medium-range modeling centers on cooperative model research and development. This study investigated TC forecast skills in the eight participating global medium-range forecast models during the year-long DIMOSIC period (June 2018 to June 2019). All models conducted 10-day forecasts from the same initial conditions based on the ECMWF IFS model cycle 45R1. The horizontal resolutions of the eight models ranged from 5 to 25km, and there were different choices of dynamical cores and physics parameterizations across the models (Magnusson et al. 2022). The forecast skills of TC track and intensity have been presented for the eight models. The TC-related precipitation and the performance of TC genesis forecasts have also been evaluated.

Comparing the model forecasts to the observations for the 109 TCs in the DIMOSIC period, IFS (45R1) and the updated version IFS-47R3 shows the best global averaged TC track forecasts, followed by ICON and SHiELD. CMC shows a relatively higher error than others before the 72-hour lead time. Based on our preliminary investigates, it could be related to the initializing moisture shock given that the CMC has much moister analyses than IFS which may induce convection collapses after the initialization with IFS ICs. For the TC track forecasts in different sub-regions, UM and ICON show the lowest track errors in the NATL, while SHiELD had the best track forecasts in the EPAC. In the WPAC and the SHEM, both IFS-47R3 and ICON show the lowest TC track errors, followed by SHiELD. From the analyses of along-track (AT) and cross-track (CT) errors, the models behave differently in different sub-regions. All models showed a slow-moving bias and a poleward bias in the NATL and the WPAC, but in the EPAC, except for JMA, most models show a fast-moving bias. In contrast, there are no consistently slow- or fast-moving biases among models in the SHEM.

For TC intensity forecasts, based on the TC tracker results using the interpolated 0.5 degree resolution data, SHiELD performs relatively better than other models, followed by UM and ARPEGE. From Table 2, we can see that the resolutions of the models range between 5 and 25 km, and the resolution of SHiELD is in the middle of that range. Therefore, the outperforming TC intensity by SHiELD may imply that the resolution is not the only major factor limiting TC intensity in global models. The use of dynamics and physics in the model also plays important role. The performance of TC track and intensity forecasts could reveal some of the characteristics of a model especially related to its dynamics and physics interactions. In Chen et al. (2019b), it has been demonstrated that updating the GFS dynamical core to the nonhydrostatic FV3 (Lin 2004; Putman and Lin 2007; Harris et al. 2020) can largely improve TC intensity forecasts, and additional improvements in TC intensity and genesis forecasts were seen when replacing the Zhao-Carr cloud microphysics scheme with the advanced GFDL cloud microphysics scheme (Zhou et al. 2019).

Here, we attempt to probe into the characteristics of the models based on their biases of TC track and intensity. Figure 12 shows the scatter plots of TC track and intensity errors for all of the forecasts during the 72-120-hour lead time in each model. Some similarities can be found in the scattered distributions of the two IFSs and ICON, including both ranges of intensity bias and track error. This is consistent with the findings in Magnusson et al. (2022) that IFS and ICON behave relatively similarly due to the sharing of partial physical parameterizations sharing between ECWMF and DWD. In contrast, SHiELD, ARPEGE, and JMA show rather unique patterns their own.

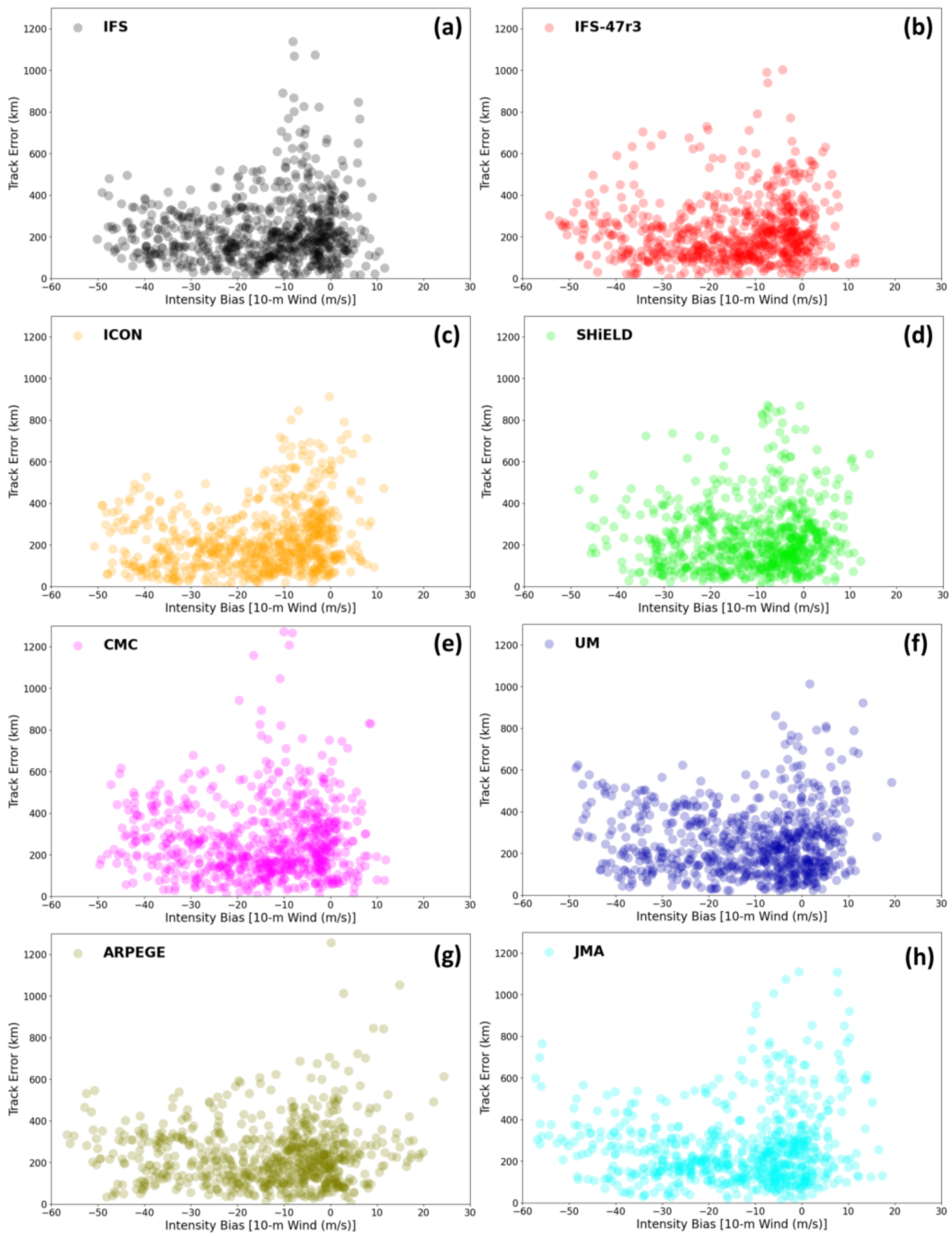

FIGURE 12. Scatter plot distribution of track errors (unit: km; ordinate) and intensity biases (the maximum 10-m wind speed; unit: m s^{-1}; abscissa) of all forecasts during the lead time of 72-120 hour for (a) IFS, (b) IFS-47r3, (c) ICON, (d) SHiELD, (e) CMC, (f) UM, (g) ARPEGE, and (h) JMA.

Figure 12 also shows that in most models, forecasts with larger track error (>700 km) are usually accompanied by smaller intensity biases (<10 ms^{-1}). In contrast, for those forecasts with larger intensity biases, their track errors are not consistently larger. At GFDL, it has been noticed that when the performance of TC track forecasts was improved by using an advection scheme with a stronger damping in the dynamics, a degradation of TC intensity was observed. 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 (Gao et al. 2021). This was attributed to the impact of stronger damping, which suppresses finer-scale features and activities, e.g. grid-scale convection in the TCs, which further suppresses the TC intensities. An in-depth study of the impact of grid-scale convection activity on TC track forecasts in SHiELD is in preparation.

Since the interpolated data cannot fully represent the performance of TC intensity forecasts in the models, the TC-related precipitation was also evaluated to provide another perspective on forecasted TC characteristics. Compared to the GPM observation data, all models under-predict the amount of TC-related precipitation, especially in the Pacific Ocean. UM better captures the regions with annual accumulated TC-related precipitation of more than 300mm compared to the other models. However, when considering all TC-related precipitation areas, the ETS of UM is generally lower than that of other models. This could be related to the relatively large track errors of UM, especially in the EPAC. In contrast, SHiELD shows the largest dry bias in TC-related precipitation among all models, but it still achieves relatively high ETSs in the EPAC due to its better track forecasts in this sub-region. As to the intensity of TC-related precipitation, most models over-predict medium precipitation events but under-predict light and heavy precipitation events. Among all models, SHiELD noticeably over-predicts precipitation events at the intensity of 1-5 mm $(6\ h)^{-1}$. The SHiELD development team at GFDL will take a close look at its low precipitation amount and over-predicted medium intensity precipitation events in the future.

The assessment of TC genesis forecast skill was based here on hits and misses, measures that showed significant inter-model variability across the different sub-regions of interest. All models show the highest hit ratios with the fewest total genesis events in the EPAC, which indicates that models can better predict TC genesis in the EPAC. In contrast, models generate more false alarms in the SHEM. SHiELD shows the highest hit ratios globally, followed by the two IFSs and ICON. CMC also achieves high hit ratios in the EPAC. In contrast, UM shows lower hit ratios than other models. As for the missed TC genesis cases, JMA missed 30 of the 100 observed TC geneses during the target period, which is the most among all models. Note that JMA uses the coarsest model resolution among all participating models, which could impair its TC genesis performance. In contrast, SHiELD shows the least number of missed TCs globally, which may be benefiting from its better TC intensity forecast.

We also investigate how well the participating models predict the timing of TC genesis by comparing the "length of lead time" proposed by Chen et al. (2019b). The results show that models can generally accurately predict TC formation earlier in the EPAC than in other sub-regions, except for JMA which predicts the WPAC and the SHEM hit events earlier than in other sub-regions. SHiELD generally predicts the earliest hit events globally, while ARPEGE also predicts TC genesis events earlier than other models in the NATL and the SHEM. Based on the differences between the model genesis lead time and the observed genesis lead time, IFS shows the most accurate timing of the TC genesis forecast, followed by JMA and ARPEGE. In contrast, the accuracy of genesis timing in ICON is relatively lower than that of other models. We also found that most models develop TCs earlier than observed in the best track, except for JMA. In addition, more than 89% of hit events in each of the models occur within 48 hours of the observed genesis time.

The comparison between IFS (version 45R1) and IFS-47R3 provides an opportunity to examine the incremental change obtained for an upgrade of one model. The upgrade from version 45R1 to 47R3 includes many changes in data assimilation and model physics. The changes and meteorological impacts have been documented by ECMWF on the website: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model. One of the listed impacts from the upgrade is the improvement of TC position errors. From our analysis, the average track errors during the first 120 hours of IFS-47R3 are 2.5 to 10.8 km less than IFS (Fig. 3) in the four major sub-regions, which is consistent with the ECMWF implementation report. However, from the analyses of the along-track and cross-track errors, the biases of slow and poleward movement are similar in these two model versions. Possibly associated with the major upgrade to moist physics (Bechtold et al. 2020), IFS-47R3 shows a slightly larger negative TC intensity biases than the older IFS version which was also found in Magnusson et al. (2021). Although total precipitation predictions remain similar across the two IFS versions, the newer IFS-47R3 achieves higher ETS scores for large accumulations, especially in the WPAC and the SHEM. We also found that IFS-47R3 has more precipitation events with stronger precipitation intensity (10-50 mm $(6\ h)^{-1}$) than IFS. However, as to the TC genesis forecast, IFS-47R3 shows some degradation from IFS, including more missed TC genesis event, higher miss ratios, shorter genesis lead times, and less accuracy on TC genesis timing. These degradations in TC genesis performance may be related to the weaker TC intensities in the newer version.

To summarize, in this study, extensive evaluation was made of the performance of TC forecasts in the DIMOSIC models based on one year of model predictions initialized with the same initial conditions. Although it is hard to precisely isolate the influence of individual components in different model formulations on TC forecast skill in such an overview, the comparisons based on different evaluation metrics highlight important similarities and differences between the models. The results will be valuable for model developers in participating centers as a benchmark of TC forecast skill with the impact of the initial condition quality removed. Also, common forecast biases of the TC movement and TC-related precipitations indicate general deficiencies in DIMOSIC models and point out a direction for model developers for further model improvement.

Acknowledgments

The authors thank Kun Gao, Jie Chen, Morris Bender, and Tom Knutson for GFDL internal review, and Lucas Harris, James Doyle, and Simon Lang for their comments helped to improve this article. Authors also would like to thank other DIMOSIC participants, Duncan Ackerley, Yves Bouteloup, K. C. Kwon, Yoonjin Lim, Mio Mastueda, Takumi Matsunobu, and Yamaguchi Munehiko for their contribution to the DIMOSIC project.

Open Research

All DIMOSIC model interpolated data can be requested from Linus Magnusson. All TC analyses are archived in the GFDL Tape Archive System at /archive/jhc/DIMOSIC/Analysis/TC and can be requested from Jan-Huey Chen.

References

Becthold, P., R. Forbes, I. Sandu, S. Lang, & M. Ahlgrimm (2020), A major moist physics upgrade for the IFS. *ECMWF Newsletter, 164*, URL: https://www.ecmwf.int/en/newsletter/164/meteorology/major-moist-physics-upgrade-ifs

Chen, J.-H., S.-J. Lin, L. Magnusson, M. A. Bender, X. Chen, L. Zhou, B. Xiang, S. L. Rees, M. J. Morin, & L. Harris (2019a), Advancements in hurricane prediction with NOAA's next generation forecast system. *Geophysical Research Letters, 46* (8), 44954501, doi:10.1029/2019GL082410.

Chen, J.-H., Lin, S., Zhou, L., Chen, X., Rees, S., Bender, M., & Morin, M. (2019b), Evaluation of Tropical Cyclone Forecasts in the Next Generation Global Prediction System, *Monthly Weather Review, 147*(9), 3409-3428, doi: 10.1175/MWR-D-18-0227.1

DWD (2022), ICON : Icosahedral Nonhydrostatic Weather and Climate Model. *Technical Report*, *DWD*. URL: https://code.mpimet.mpg.de/projects/iconpublic/wiki/Documentation

ECMWF (2018), IFS Documentation CY45R1. *Technical Report*, *ECMWF*. URL: https://www.ecmwf.int/en/publications/ifs-documentation

ECMWF (2021), IFS Documentation CY47R3. *Technical Report*, *ECMWF*. URL: https://www.ecmwf.int/en/publications/ifs-documentation

Gao, K., L. Harris, L. Zhou, M. A Bender, and M. Morin (2021), On the sensitivity of hurricane intensity and structure to horizontal tracer advection schemes in FV3. *Journal of the Atmospheric Sciences*, 78(9), doi:10.1175/JAS-D-20-0331.13007-3021.

Girard, C., & Coauthors (2014), Staggered vertical discretization of the Canadian Environmental Multiscale (GEM) model using a coordinate of the log-hydrostatic-pressure type. *Monthly Weather Review, 142* (3), 1183.

Harris, L. M., S.-J. Lin, & C. Tu (2016), High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *Journal of Climate, 29*, 4293–4314, doi:10.1175/JCLI-D-15-0389.11196, doi:10.1175/MWR-D-13-00255.1.

Harris, L., & Coauthors (2020), GFDL SHiELD: A unified system for weather to seasonal prediction. *Journal of Advances in Modeling Earth Systems*, *12* (10), doi:10.1029/2020MS002223.

Hong, Y., K.-L. Hsu, S. Sorooshian, & X. Gao (2004), Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *Journal of Applied Meteorology*, *43* (12), 18341853, doi:10.1175/JAM2173.1.

JMA (2019), Outline of the operational numerical weather prediction at the Japan Meteorological Agency, *Technical Report, Japan Meteorological Agency*. URL http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2019-nwp/index.htm.

Lin, S.-J. (2004), A "vertically Lagrangian" finite-volume dynamical core for global models, *Monthly Weather Review*, 132, 2293-2307.

Magnusson, L., J.-H. Chen, S.-J. Lin, L. Zhou, & X. Chen (2019), Dependence on initial conditions vs. model formulations for medium-range forecast error variations. *Quarterly Journal of the Royal Meteorological Society*, *145*, doi:10.1002/qj.3545

Magnusson, L., & co-authors (2021), Tropical cyclone activities at ECMWF. *ECMWF Technical Memorandum 888*

Magnusson, L., D. Ackerley, Y. Bouteloup, J.-H. Chen, J. Doyle, P. Earnshaw, Y. C. Kwon, M. Köhler, S. T. K Lang, Y.-J. Lim, M. Matsueda, T. Matsunobu, R. McTaggart-Cowan, A. Reinecke, M. Yamaguchi1, & L. Zhou (2022), Skill of medium-range forecast models using the same initial conditions. *Bulletin of the American Meteorological Society*, *103*(9), E2050-E2068, doi:10.1175/BAMS-D-21-0234.1

McTaggartCowan, R., & Coauthors (2019), Modernization of atmospheric physics parameterization in Canadian NWP. *Journal of Advances in Modeling Earth Systems, 11* (11), 35933635, doi:10.1029/2019MS001781.

Miller, R. J., A. J. Scrader, C. R. Sampson, & T. L. Tsui (1990), The Automated Tropical Cyclone Forecast System (ATCF). *Weather and Forecasting*, *5*, 653–660.

Mogensen, K. S. , L. Magnusson , & J.-R. Bidlot (2017), Tropical cyclone sensitivity to ocean coupling in the ECMWF coupled model: Tropical cyclone sensitivity. *J. Geophys. Res. Oceans*, *122*, 4392–4412, doi:10.1002/2017JC012753.

Pollard, R. T. , P. B. Rhines , & R. O. R. Y. Thompson (1973), The deepening of the wind-mixed layer. *Geophys. Fluid Dyn.* , *4*, 381–404, doi:10.1080/03091927208236105.

Putman, W. M., & S.-J. Lin (2007) Finite-volume transport on various cubed-sphere grids. *Journal of computational physics*, 227, 55-78, doi:10.1016/j.jcp.2007.07.022.

Rabier, F., H. Ja¨rvinen, E. Klinker, J.-F. Mahfouf, & A. Simmons (2000), The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society, 126* (564), 24 1143–1170, doi:10.1002/qj.49712656415

Roehrig, R., & Coauthors (2020), The cnrm global atmosphere model arpegeclimat 6.3: Description and evaluation. *Journal of Advances in Modeling Earth Systems*, *12* (7), doi:10.1029/2020MS002075.

Schaefer, J. T. (1990) The critical success index as an indicator of warning skill. *Weather and Forecasting, 5* , 570–575.

Sampson, C. R., & A. J. Schrader (2000), The Automated Tropical Cyclone Forecasting System (version 3.2). *Bulletin of the American Meteorological Society*, *81*, 1231–1240.

Walters, D., & Coauthors (2019), The met office unified model global atmosphere 7.0/7.1 and jules global land 7.0 configurations. *Geoscientific Model Development*, *12*(5), 19091963, doi:10.5194/gmd-12-1909-2019.

Zeng, X. , & A. Beljaars (2005), A prognostic scheme of sea surface skin temperature for modeling and data assimilation: Sea surface skin temperature scheme. *Geophysical Research Letters, 32*, L14605, doi:10.1029/2005GL023030.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, & S. Rees (2019), Toward convective scale prediction within the Next Generation Global Prediction System. *Bulletin of the American Meteorological Society*. 100 (7): 1225-43, doi:10.1175/BAMS-D-17-0246.1

Tropical Cyclone Forecasts in the DIMOSIC Project -- Medium-Range Forecast Models with Common Initial Conditions

Jan-Huey Chen[1,2], Linjiong Zhou[3], Linus Magnusson[4], Ron McTaggart-Cowan[5], and Martin Köhler[6]

[1]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

[2]University Corporation for Atmospheric Research, Boulder, CO, USA

[3]Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, NJ, USA

[4]European Centre for Medium-Range Weather Forecasts, Reading, UK

[5]Environment and Climate Change Canada, Montreal, Canada

[6]Deutsche Wetterdienst, Offenbach, Germany

Corresponding author: Jan-Huey Chen (Jan-Huey.Chen@noaa.gov)

Key Points:

- Tropical cyclone forecasts are compared between global medium-range models from leading modeling centers initialized with identical data.
- Similarities and differences between the models set a benchmark of TC forecast with the impact of the initial condition quality removed.
- Common TC forecast biases indicate general deficiencies in the models and suggest a direction for further model improvement.

Abstract

The Tropical cyclone (TC) forecast skill of the eight global medium-range forecast models which are participating in the DIMOSIC (DIfferent Models, Same Initial Conditions) project is investigated in this study. Each model was used to generate 10-day forecasts from the same initial conditions provided by the European Centre for Medium-Range Weather Forecasts. There are a total of 123 initial dates spanning in one year from June 2018 to June 2019 with a 3-day interval. The TC track and intensity forecasts are evaluated against the best track dataset. TC-related precipitation and tropical cyclogenesis forecasts are also compared to explore the differences and similarities of TC forecasts across the models. This comparison of TC forecasts allows model developers in different centers to benchmark their model against other models, with the impact of the initial condition quality removed. The verifications reveal that most models show slow-moving and right-of-track biases in their TC track forecasts. Also, a common dry bias in TC-related precipitation indicates a general deficiency in TC intensity and convection in the models which should be related to insufficient model resolution. These findings provide important references for future model developments.

Plain Language Summary

Despite recent improvements in our ability to predict the track and intensity of tropical cyclones, these storms remain significant forecasting challenges. Forecasters rely heavily on the guidance generated by numerical weather prediction systems, making the reliability of these systems essential for accurate forecasts during these high-impact weather events. As a result, improvement the quality of tropical cyclone guidance is an important numerical model development objective. In this study, the TC forecast skills in the eight global medium-range forecast models from the model development centers/institutes who participated in the DIMOSIC (DIfferent Models, Same Initial Conditions) project are examined. All models were initialized from the same data provided by the ECMWF (European Centre for Medium-Range Weather Forecasts) to investigate the differences and similarities among their TC forecasts without the impact of the quality of initial conditions. Besides the general TC forecast evaluation metrics including errors and biases of the track and intensity, the TC-related precipitation and TC genesis skills are also evaluated to comprehensively explore the performance of TC forecasts among all models. The comparison allows model developers in different centers to benchmark their model against other participating models. Moreover, the verification results provide important references for future model developments.

1 Introduction

Tropical cyclone (TC) prediction is an important mission for weather and climate agencies in many countries. Over the past few decades, numerical models have become the most important tools for operational centers to make TC forecasts on weather and sub-seasonal to seasonal time scales. Therefore, improving the model performance of TC forecasts has been one of the leading tasks in most operational centers or modeling research institutes working on model development. In addition, the accurate depiction of physical processes that lead to a better TC forecast in the model are also relevant to interesting scientific questions in the atmospheric science research area more broadly.

The quality of initial conditions has a leading impact on short- to medium-range forecast skill, including for TC forecasts. In Chen et al. (2019a), the fvGFS (finite volume Global Forecasting System) model developed at the Geophysical Fluid Dynamics Laboratory (GFDL) initialized with the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) data showed much-improved TC track forecasts for the 2017 Atlantic hurricane season compared to its retrospective forecasts initialized with the data from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) version 14. In Magnusson et al. (2019), the same approach was used, comparing the GFDL fvGFS model forecasts to those from the IFS and GFS. The results showed that the choice of initial conditions clearly dominated the forecast quality in the medium-range predictions, but that the model formulation could also play a significant role.

Since major model development centers mostly develop their modeling systems independently, the DIMOSIC project (DIfferent Models, Same Initial Conditions; Magnusson et al. 2022) was devised to investigate the relationship between the choice of model formulation and forecast quality. Models developed by different world-leading modeling centers were initialized from the same initial condition. In Magnusson et al. (2022), the differences and similarities of the forecasts among the models were presented. The results found that some pairs of models behaved more similarly than other pairs due to their sharing of partial physical parameterizations, e.g. ECWMF IFS and DWD (Deutsche Wetterdienst) ICON (Icosahedral Non-hydrostatic Model). On the other hand, ICON and GFDL SHiELD (System for High-Resolution Prediction on Earth-to-Local Domains) showed relatively large forecast differences, while both ranking among the best models. Regarding the influences from model formulations on the forecasts, however, it was difficult to point out a single model component that had the strongest impact on the forecast differences. Also, as pointed out by Magnusson et al. (2022) the interaction between different model parameterizations and their respective configurations could play a significant role as well.

In this study, the performance of TC forecasts from the DIMOSIC models is evaluated. The TC track and intensity forecast skills among the models during the period of June 2018 to June 2019 are compared. Since TC intensity in interpolated data does not reflect the actual TC intensity at the native model resolution, the TC-related precipitation are also evaluated to provide another perspective on forecasted TC activities for better exploring the differences and similarities among the models. Also, the forecast skill of TC genesis was investigated by comparing the hit/false alarm ratios among the models, as well as using the method based on the lengths of TC genesis lead time introduced in Chen et al. (2019b) to examine the accuracy of TC genesis timing in the model forecasts. These comparisons should be valuable for model developers in different centers to benchmark their model's performance on TC forecasts against that of other models, with the impact of the initial condition quality removed.

The models participating in the DIMOSIC project are introduced in section 2 which also describes the observation data and methodology used in this study. The comparisons of track, intensity, TC-related precipitation, and genesis forecasts among the models are contained in section 3. Summary and discussion are presented in section 4.

2 DIMOSIC models, forecasts, and verification data

General information on the numerical models and their developing centers/institutes participating in the DIMOSIC project are listed in Table 1. The horizontal resolutions and the number of vertical levels of the models and their key references are included. Some centers/institutes submitted more than one model configurations to the project, but we only investigate one configuration of the model for each center/institute based on their suggestions. The only exception is to include two versions of IFS 45R1 and 47R3, to provide an example of the incremental change obtained for an upgrade of one model. For the sea surface temperature evolution in the models, the two IFSs used a partial coupling to the 3D ocean NEMO model (Mogensen et al. 2017), SHiELD is coupled with a 1D mixed layer ocean model (Pollard et al. 1973), CMC used a thermodynamic mixed layer ocean model (Zeng and Beljaars 2005), and others used persistent anomalies from the analysis. Other detailed configurations of each model including dynamical cores and major physical parameterizations can be found in the sub-section of “Model descriptions” in the section of “Models and data“ and in Table 2 in Magnusson et al. (2022), but is not repeated in this paper.

Acronyms	Models	Centers/Institutes	Resolution	Key references
ARPEGE	Action de Recherche Petite Echelle Grande Echelle (version: 46T1)	Meteofrance	5-25 km 105 levels	Roehrig et al. (2020)
CMC	Global Environmental Multiscale Model (GEM) (version: v5.0.2)	Canadian Meteorological Center (CMC)	15 km 80 levels	Girard et al. (2014) McTaggart-Cowan et al. (2019)
ICON	Icosahedral Non-hydrostatic Model (version: April 21)	Deutsche Wetterdienst (DWD)	13km 90 levels	DWD (2022)
IFS/ IFS-47R3	Integrated Forecasting System (versions: 45R1 and 47R3)	European Centre for Medium-range Weather Forecasts (ECMWF)	9 km 137 levels	ECMWF (2018, 2021)
JMA	Global Spectral Model (GSM) (version: GSM1705)	Japan Meteorological Agency (JMA)	20 km 100 levels	JMA (2019)
SHiELD	System for High-Resolution Prediction on Earth-to-Local Domains (version: rt2019)	Geophysical Fluid Dynamics Laboratory (GFDL)	13 km 91 levels	Harris et al. (2020)
UM	Unified Model	UK Met Office	10 km 70 levels	Walters et al. (2019)

TABLE 1. The eight DIMOSIC models (the ECWMF contributed two IFS configurations). From left to right: model acronyms, model full names (and versions if applicable), centers/institutes that models belong to, horizontal resolutions and the number of vertical levels used in the models, and the key references of the models.

All models conducted 10-day forecasts from the same initial conditions: ECMWF operational analyses based on the IFS model cycle 45R1 (ECMWF 2018). The 9-km analyses on 137 vertical levels are generated from a 4DVar data assimilation system (Rabier et al. 2000). All participating institutes received the interpolated data at 0.1 degree for their model initialization. Detailed procedures for handling the initial conditions in each model are described in section 2c and Table 3 of Magnusson et al. (2022).

The total 123 forecasts are conducted with initialization dates spanning one year from June 2018 to June 2019 at 3-day intervals. The 10-day forecast outputs from each model were interpolated to a common 0.5 degree grid using an average interpolation method available in the EcCodes/MIR package (https://confluence.ecmwf.int/display/ECC/ecCodes+Home). The GFDL simpler tracker (Harris et al. 2016) was used with a warm-core criterion to track TCs in the forecasts of the eight models based on the fields of sea-level pressure, 10-m wind speed, 850-hPa vorticity, and mean temperature between 500-300 hPa.

There were 109 observed TCs in the DIMOSIC period. Their storm tracks based on the best track data (b-deck) in Automated Tropical Cyclone Forecast (ATCF) dataset (Miller et al. 1990; Sampson and Schrader 2000) are shown on the map in Fig 1. There were 35 TCs in the northwest Pacific basin (WPAC), 24 in the northeast Pacific basin (EPAC), and 16 in the North Atlantic basin (NATL). In the Southern Hemisphere (SHEM), there were 27 TCs in the combined South Indian Ocean and South Pacific Ocean. Besides the overview of global analyses, the individual TC forecast skill in the above four major sub-regions will be investigated and compared with each other in the following sections.

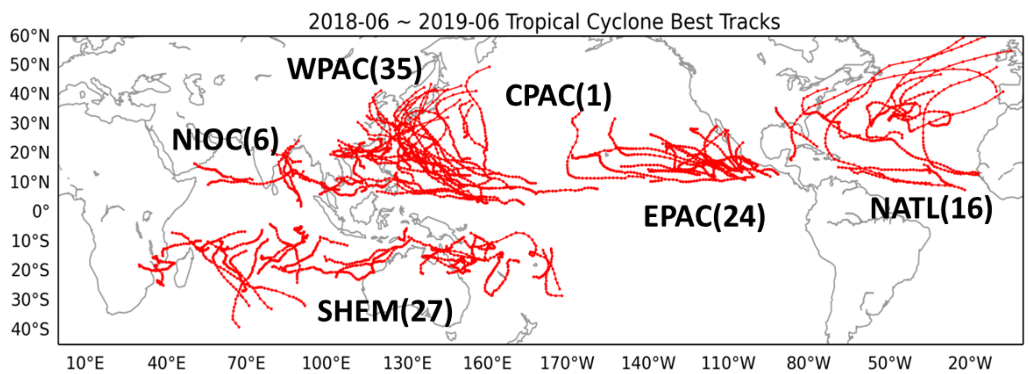

FIGURE 1. All TCs in the DIMOSIC period. The best tracks are from the ATCF dataset. The numbers of TCs are indicated in the brackets next to the acronyms of the six regions: WPAC: northwest Pacific basin; EPAC: northeast Pacific basin; CPAC: north-central Pacific basin; NATL: North Atlantic basin; NIOC: north Indian Ocean; SHEM: Southern Hemisphere.

The ATCF dataset was used to evaluate the forecast errors of TC track and intensity, and the skill of TC genesis forecasts in the models at six-hour intervals. For the TC-related precipitation, the NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) dataset was used (Hong et al. 2004) for verification. To equally compare to the model output field of total precipitation, this high-resolution (0.1 degree) satellite observational dataset was interpolated into 0.5 degree.

3. Results

3.1 TC track forecasts

The prediction of TC path, or track, is the one of the most important factors for taking necessary precautions against possible impacts from hurricanes or typhoons. The homogeneous comparisons of global mean TC track forecast errors along with the forecast lead time at 12-h intervals are shown in Fig. 2a. The differences among the models are small through the 72-hour lead time with the exception of CMC which shows a relatively higher error than others. During the 72 to 120-hour head time, TC track errors diverge into two groups. The two IFSs, ICON, SHiELD, and ARPEGE show lower errors than UM, CMC, and JMA. Both versions of IFS show the lowest TC track errors after the 120-hour lead time all the way to the 168-hour lead time, followed by ICON and SHiELD. Note that most models are within the 95% confidence levels of IFS, which means the differences in forecast skills are not statistically significant. To highlight the difference between the leading models, Fig. 2b shows the differences in TC track errors of five of the models compared to the IFS. The newer IFS-47R43 performs slightly better (negative values in track error differences) or equivalently during the entire 7 days. In the early lead times (36-84 hours), most of the five models also show slightly better forecasts than IFS. ICON displays similar (or slightly better) skill to the two IFS versions until the 120-hour lead time. Both SHiELD and ARPEGE perform very well before the 96-hour lead time. After the 120-hour lead time, SHiELD shows lower forecast errors than other models except for the two IFSs.

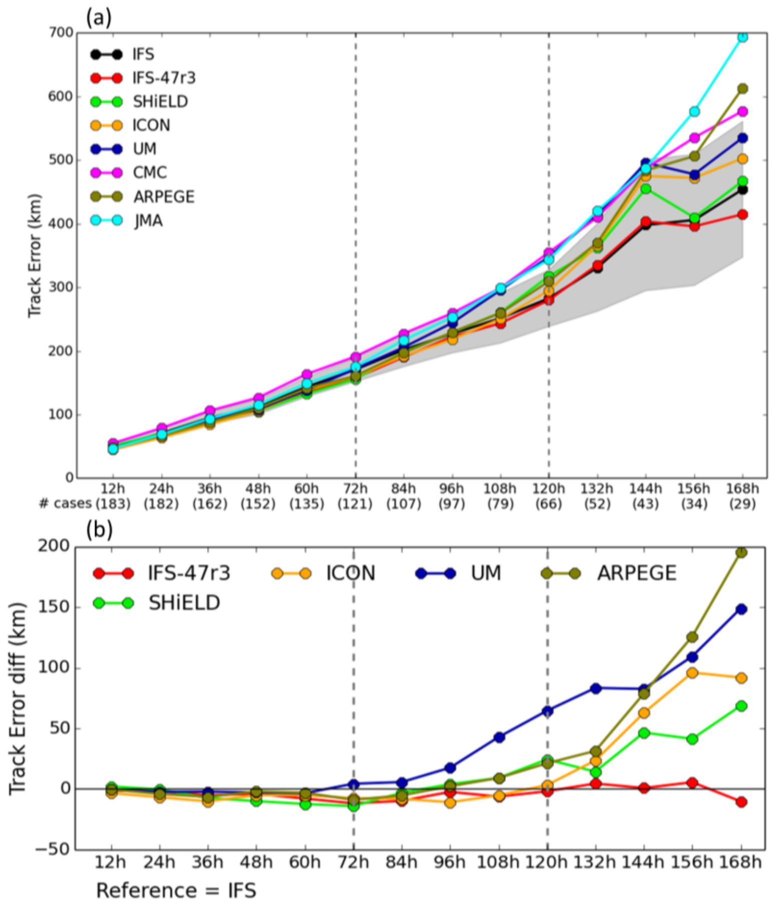

FIGURE 2. (a) Global mean TC track forecast errors (km) at every 12 hour forecast lead time for IFS (black), IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), CMC (magenta), ARPEGE (grass green), and JMA (light blue). The 95% confidence levels for IFS are indicated by the gray color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom of each abscissa. Vertical gray dotted lines indicate 72 and 120 hour forecast lead times. (b) Global mean TC track forecast error differences of IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), and ARPEGE (grass green) comparing to IFS.

Figure 3 shows the 5-day average TC track errors for all models in the entire globe and in the four major sub-regions individually. For the two IFSs, IFS-47R3 shows lower track errors than IFS globally and in all major sub-regions. ICON shows competitive low track errors to IFS-47R3 in the WPAC and the SHEM, and a very low track error in the NATL. SHiELD performs the lowest track error in the EPAC, and low track errors besides the two IFSs and ICON in the SHEM and the WPAC. However, SHiELD has a much larger track error in the NATL, which results from the slow moving bias shown in the forecasts of Hurricane Florence and the bias of direction of motion shown in the forecasts of Hurricane Leslie. Detailed investigations can be found in Text S1 and Figs. S1-S3. The performance of the TC track forecast in UM is notable.

The model shows the lowest track error in the NATL but the largest track error in the EPAC, while its track errors in the WPAC and the SHEM are also at the high end compared to other models. JMA performs competitively to IFS-47R3 and ICON in EPAC, but not in other sub-regions. With regard to CMC, it shows larger track errors than other models in most sub-regions except for in the EPAC during this targeted year. From Fig. 3, we can also see that the globally-averaged track forecast errors among the models is dominated by the errors in the WPAC and the SHEM since the majority of the TCs were found in these two sub-regions.

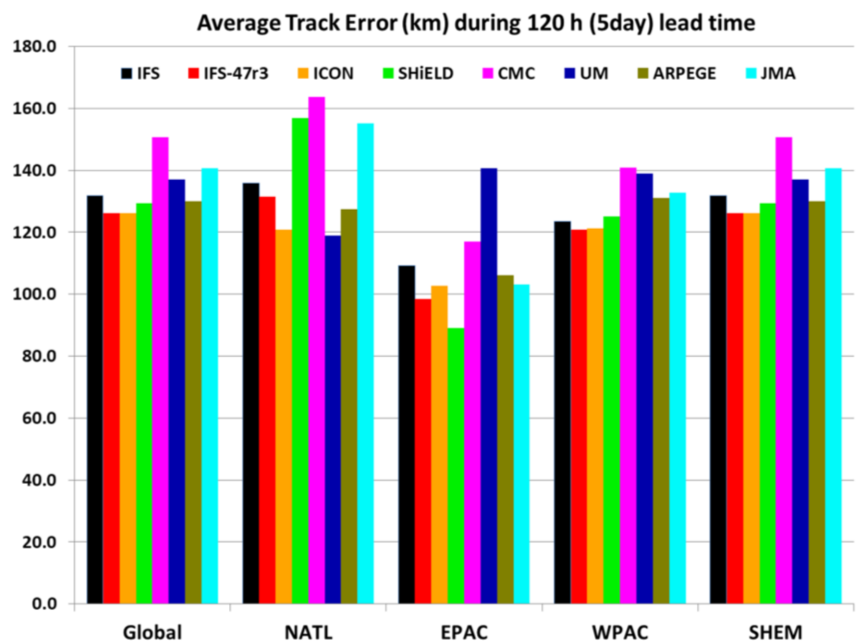

FIGURE 3 Averaged Track errors (km) in globe and 4 sub-regions during the 120-hour lead time for the 8 models. Abbreviations and colors used for the models are the same as in Fig. 2a. Abbreviations used for the sub-regions on the abscissa are the same as in Fig. 1.

The sources of track errors can be due to biases either in forecasts of the TC translational speed or the TC direction of motion. The lower sub-panels in Fig. 4 show the globally averaged along-track (AT) errors and cross-track (CT) errors (perpendicular to the track) for all models. Both AT and CT errors are calculated as great circle distances. Most of the models start to show negative AT biases and positive CT biases during 72 to 120-hour lead time, which indicates that the TC track errors during the later lead times are mostly due to the slow and northward (for an easterly moving TC) moving biases. In general, UM shows the smallest AT and CT biases among all models, and the biases of SHiELD take place at longer forecast lead times than other models.

By the Pythagorean Theorem the square of the total error equals the squares of the AT and CT errors (Chen et al. 2019a). The squares of total track errors, CT errors, and AT errors are plotted in upper sub-panels in Fig. 4 to illustrate the proportion of contributions from the AT and CT errors respectively to the total error. From Figs. 4a,b, we can see that the AT error contributes more than the CT error to the total track error in the two IFSs, which indicates that the track errors in the IFSs are dominated by the slow-moving bias (negative AT biases). The characteristic of the consistently larger AT error square than the CT error square can also be found in ICON, CMC, and UM, but the differences between their AT and CT error squares are smaller than those in the IFSs. SHiELD, ARPEGE, and JMA show relatively closer AT and CT error squares, especially SHiELD. This indicates that the contributions of the slow and northward moving biases to the total track errors are similar in these models. For ARPEGE, the CT error is the dominant track error in late lead times, while the AT error contributes more to the JMA's total TC track errors during the 120 to 144-hour lead time.

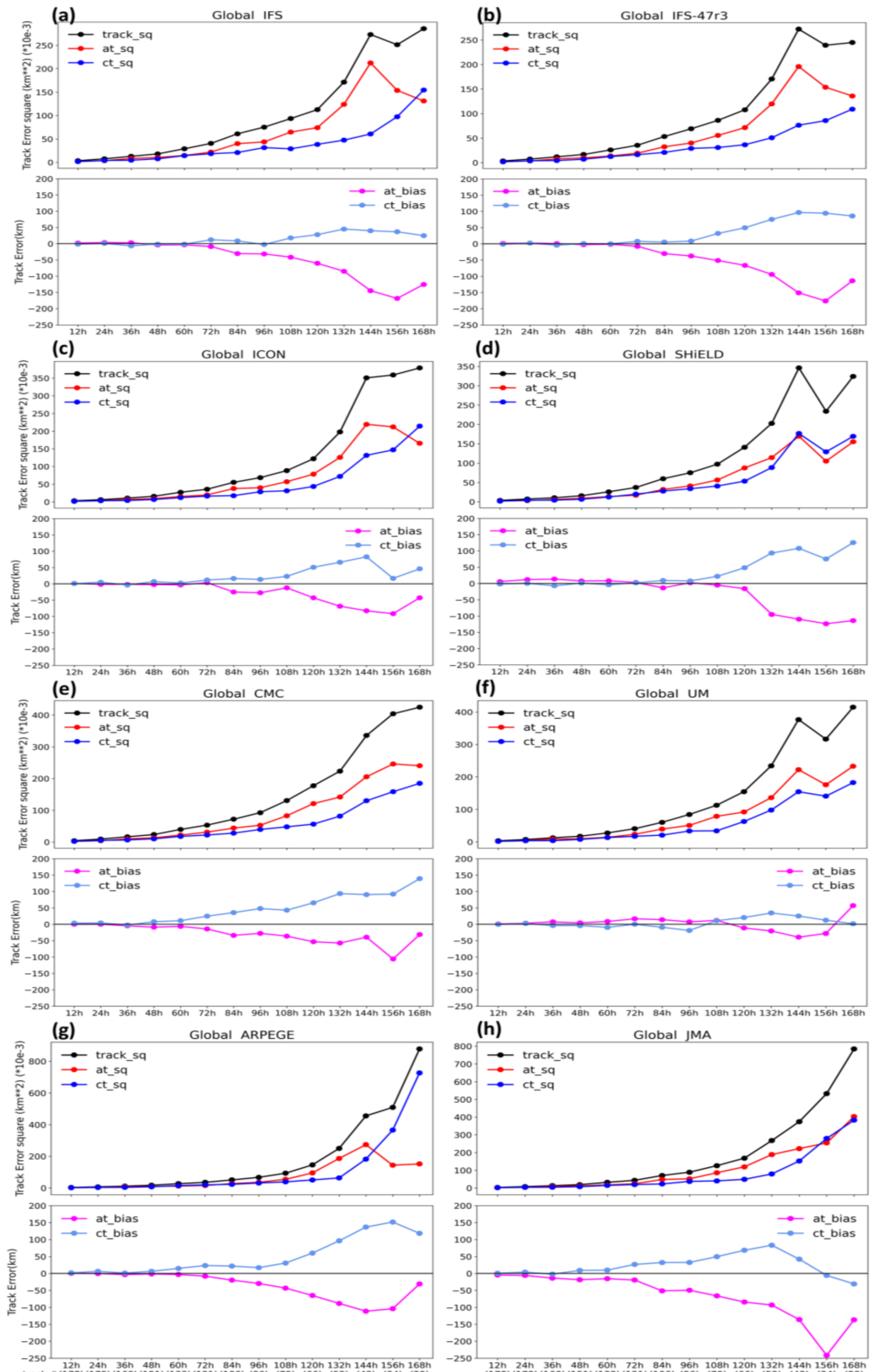

FIGURE 4 Global analyses of along-track (AT) error and cross-track (CT) error for (a) IFS, (b) IFS-47R3, (c) ICON, (d) SHiELD, (e) CMC, (f) UM, (g) ARPEGE, and (h) JMA. The squares of total track errors (black), along-track errors (red), and cross-track errors (blue) are in the upper panels for each model. The biases of along-track (magenta) and cross-track (light blue) errors are in the lower panels. Numbers of homogeneous cases for each lead time are listed at the bottom of lower panels.

It is also found that models have different AT and CT errors in different sub-regions. All models showed a slow-moving bias and a poleward bias in the NATL and the WPAC, but in the EPAC, except for JMA, most models show a fast-moving bias. In contrast, there are no consistently slow or fast moving biases among models in the SHEM. Detailed analyses of AT and CT errors for all eight models in the four major sub-regions can be found in Text S2 and Figs. S4-S7.

3.2 TC intensity forecasts

It has been more challenging to predict TC intensity than track, especially for global models which usually cannot resolve fine scale interactions between thermal dynamics and dynamics due to insufficient resolutions. As outlined in Section 2, the 10-day forecast outputs were interpolated to a common 0.5 degree for each model. Therefore, the model-predicted TC intensities found by the tracker are underestimated due to the low data resolution. However, it is still of interest to compare the relative differences of TC intensities among the models for the interpolated output data. The global mean TC intensity errors and biases based on the maximum 10-m wind speed are presented in Fig. 5a. SHiELD predicts a much stronger TC intensity than other models, followed by UM and ARPEGE. Figure 5b uses SHiELD as an example to demonstrate the differences between the TC intensities obtained from the native resolution (13 km grid) outputs and in the interpolated 0.5 degree resolution data. The average differences of total error and bias are between 3 to 5 ms^{-1}.

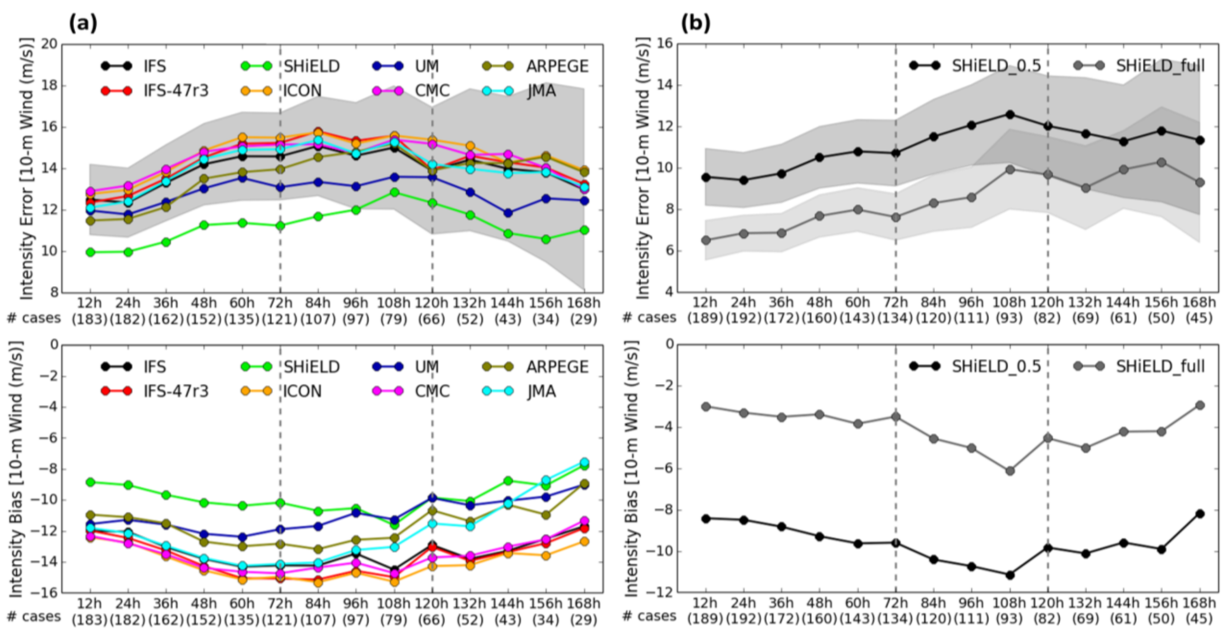

FIGURE 5. (a) Global mean TC intensity errors and biases. Upper panel: Absolute error of the maximum 10-m wind speed ($m\ s^{-1}$) along with the model forecast lead time for 8 models. Abbreviations and colors used for the models are the same as in Fig. 2a. The 95% confidence levels for IFS are indicated by the gray color shading. Numbers of homogeneous cases for individual lead times are listed in the brackets at the bottom. Vertical grey dotted lines indicate 72 hour and 120 hour lead times. Lower panel: As in the upper panel, but for the bias of the maximum 10-m wind speed ($m\ s^{-1}$). (b) As in (a), but for SHiELD native resolution data (black; SHiELD_full) and SHiELD 0.5 degree interpolated data (gray; SHiELD_0.5). The 95% confidence levels for each resolution data are indicated by the same medium and light transparent grey shading areas, with their overlapping region denoted by dark grey shading.

3.3 Forecasts of TC-related precipitation

Since the performance of TC intensity forecasts cannot be fully represented by the interpolated data, here, the TC-related precipitation is evaluated to provide another perspective on the forecasted TC characteristics in the models. Using the TC track information, the precipitation within 350 km of each TC center is used to investigate the TC-related precipitation for each model. Figure 6 shows the accumulated total precipitation for all TCs during the DIMOSIC period in each model compared to the Global Precipitation Measurement (GPM) observational data (Fig. 6a). The comparison shows that all models under-predict the amount of precipitation, especially in the most active areas of the WPAC and the EPAC. From a broad visual comparison, UM and SHiELD appear to have produced the highest and lowest amounts of precipitation among all the models, respectively. This can be confirmed by comparing the accumulated precipitation of models to the GPM data presented in Fig. 7a. The UM shows a much larger amount of precipitation than other models, followed by ARPEGE and CMC. In contrast, SHiELD shows the least TC-related precipitation among all models, except in the EPAC. The ranks of models are similar in different sub-regions, while the global precipitation amounts are dominated by those in the WPAC and the SHEM, as expected.

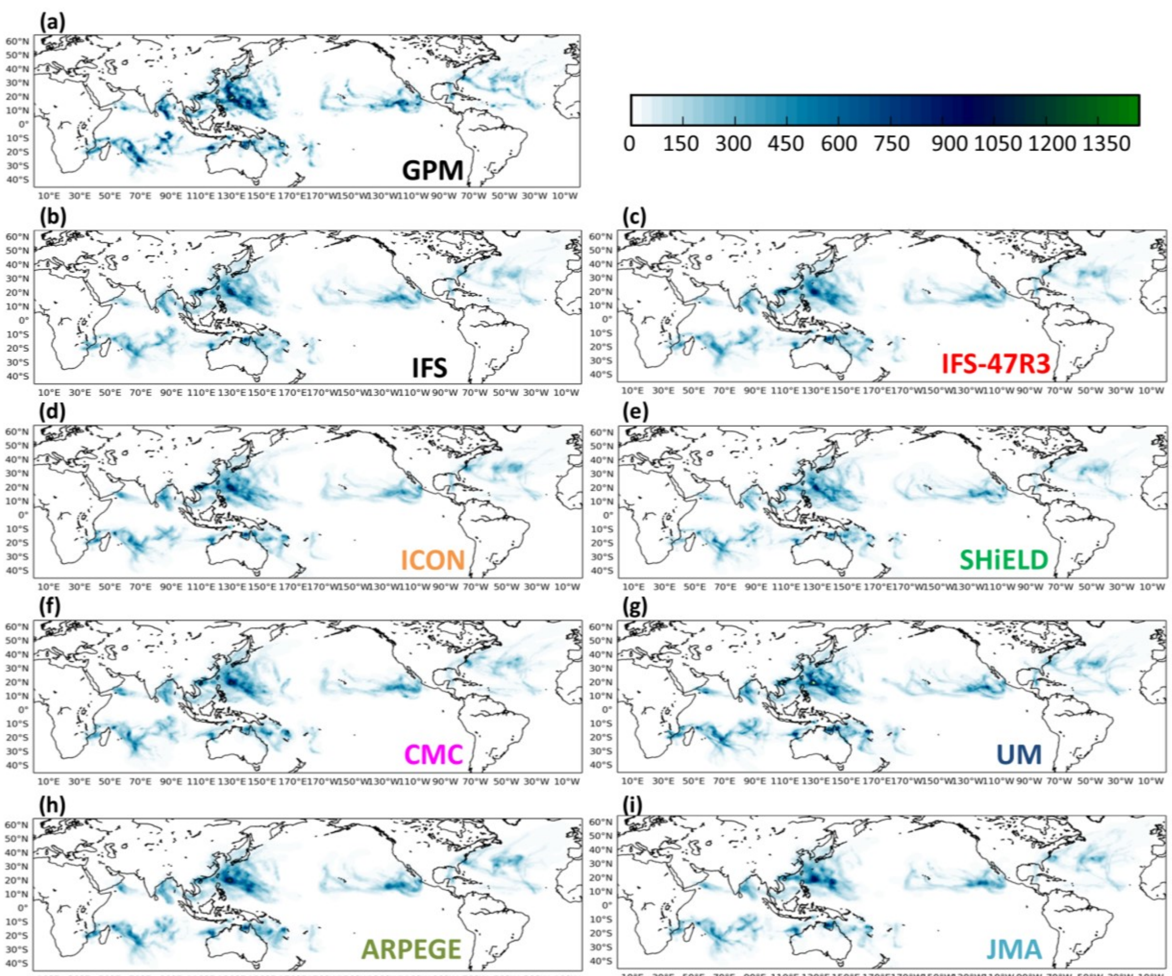

FIGURE 6. Accumulated TC-related precipitation (unit: mm) for all TCs in the DIMOSIC period in (a) Global Precipitation Measurement (GPM) observations, (b) IFS, (c) IFS-47R3, (d) ICON, (e) SHiELD, (f) CMC, (g) UM, (h) ARPEGE, and (i) JMA.

To more objectively compare the forecasted locations of TC-related precipitation in each model to the GPM observations, the equitable threat scores (ETSs; Schaefer 1990) are computed. The ETSs for all TC-related precipitation areas in the four sub-regions for all eight models are compared in Fig. 7b. Note that although UM shows the closest precipitation amount to the GPM observation data (Fig. 7a), its ETS (skill) is lower than other models when considering all TC-related precipitation areas. This could be related to its relatively larger track errors (Fig.3) that cause the displacement of precipitation locations. However, for the areas with at least 300 mm of accumulated TC-related precipitation, the ETSs of UM are generally higher than those of other models (Fig. 7c). This is likely due to its relatively better prediction of precipitation amounts (Fig. 7a). In contrast, SHiELD under-predicts the precipitation amounts, but owing to its better track forecast in the EPAC (Fig. 3), it is able to achieve relatively higher ETSs in this sub-region (Figs. 7b,c). When comparing the two IFSs, Fig. 7 shows that their accumulated precipitation amounts are similar, but the newer IFS-47R3 generally had higher ETS scores (Figs. 7b,c). Finally, JMA shows relatively higher ETSs in the WPAC in both categories, while in the NATL, the highest ETSs in the two categories are achieved by ICON and ARPEGE (Figs. 7b,c).

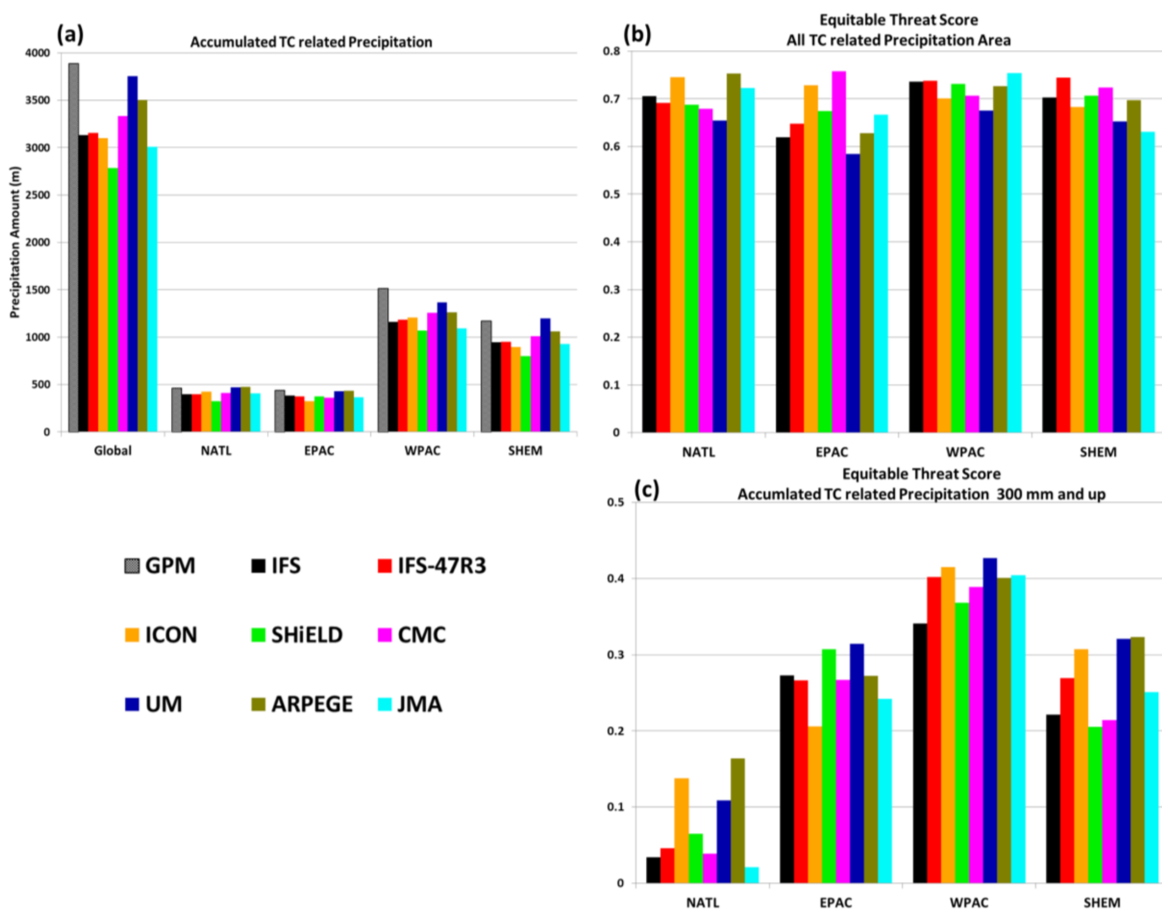

FIGURE 7. (a) Global accumulated TC-related precipitation (unit: m) and for the four sub-regions for all TCs in the DIMOSIC period. The equitable threat scores (ETSs) for (b) all TC-related precipitation areas and (c) the areas with 300 mm and up accumulated TC-related precipitation. The GPM analysis data in (a) is shown in the bars with the grey checkerboard pattern. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 3.

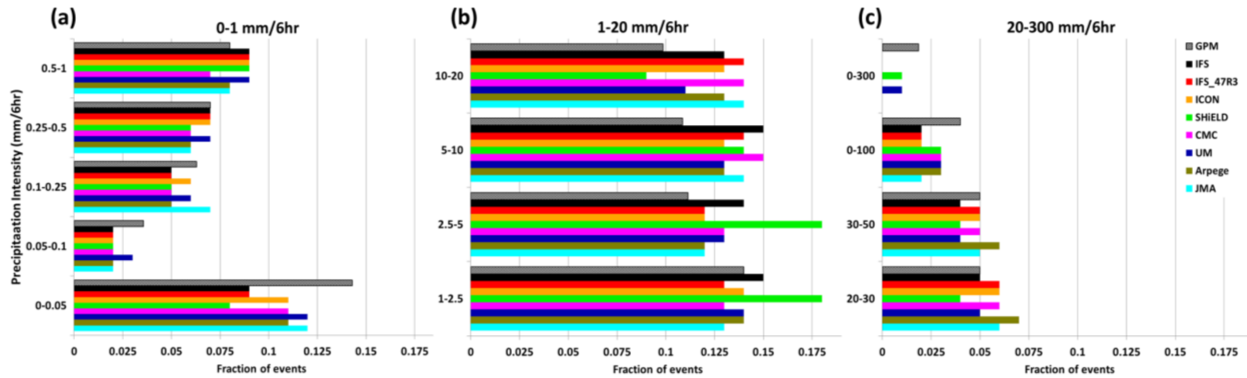

FIGURE 8. Fractions of precipitation events (GPM observations and model forecasts) in each precipitation intensity bin. (a) 5 bins for precipitation intensities from 0 to 1 mm(6 h)$^{-1}$. (b) four bins for precipitation intensities from 1 to 20 mm(6 h)$^{-1}$. (c) four bins for precipitation intensities from 20 to 300 mm(6 h)$^{-1}$. Abbreviations and colors used for the model are the same as in Fig. 7.

TC-related precipitation based on different precipitation intensities was also analyzed. Figure 8 shows the fractions of precipitation events in different precipitation intensity bins. Most of the models under-predicted light (weaker than 0.5 mm (6 h)$^{-1}$; Fig. 8a) and heavy (stronger than 50 mm (6 h)$^{-1}$; Fig. 8c) precipitations, but over-predicted medium precipitation events (Fig. 8b). Although SHiELD and UM were able to predicts some heavy precipitation events in the bin of 100-300 mm (6 h)$^{-1}$, SHiELD significantly over-predicted the events between 1-5 mm (6 h)$^{-1}$. The SHiELD development team at GFDL will closely examine the precipitation forecasts in the model in the near future, particularly to better isolate the possible reasons for these excessive precipitation amounts.

3.4 Forecasts of TC genesis

When a timeline contains an observed TC genesis, the track and intensity forecasts of the TC are verified based on the model forecasts initialized at or after the observed TC genesis time. In contrast, to investigate the models' performance for TC genesis, the 10-day forecast runs initialized before the observed TC genesis time which is based on the first "TD (tropical depression)" recorded in the ATCF best track data are considered. All TCs found by the GFDL simple tracker in these forecasts but not existing as TCs in the initial conditions for the forecast are counted as genesis events in the models. If a TC genesis has a track that "matches" an observed TC track, the genesis case is categorized as a "hit event". Otherwise, it is a "false alarm". The same criteria is used as in Chen et al. (2019b) to judge a model storm was a successful prediction of an observed genesis event.

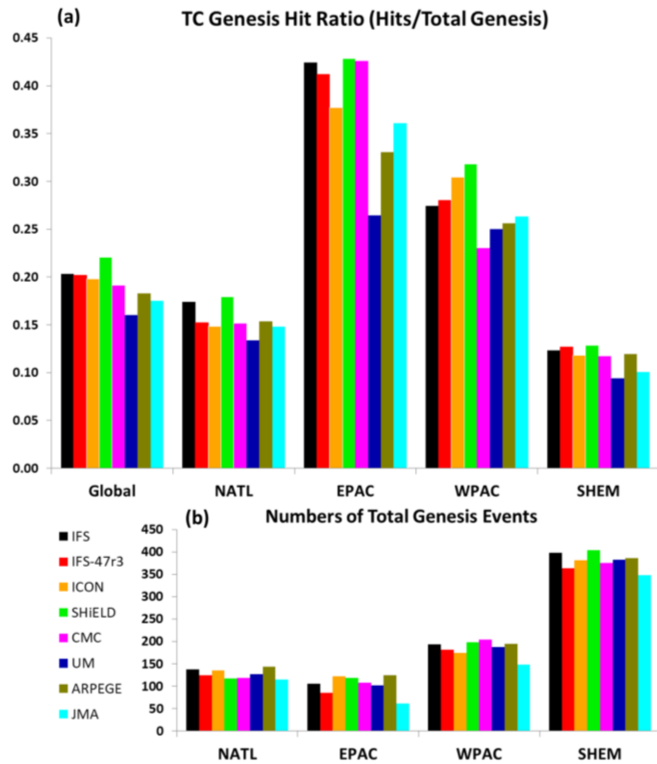

FIGURE 9. (a) Ratios of hit events to the total number of genesis events and (b) Numbers of total genesis evens (sum of hit events and false alarms) for all models. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7.

Figure 9 shows the TC genesis ratios (hits to total predicted genesis events) and the number of total genesis events for each of the eight models in different regions. The sum of hit events and false alarms is equal to the number of total forecasted genesis events. We first find that all models show the highest hit ratios (Fig. 9a) with the fewest total genesis events in the EPAC. This indicates that models can predict TC genesis more skillfully in the EPAC than in other sub-regions. In contrast, models show the lowest hit ratios but the largest numbers of total genesis events in the SHEM, which indicates that models generate more false alarms in this sub-region than in the others. In general, SHiELD demonstrates the highest hit ratios both globally and in all sub-regions, followed by the two IFSs and ICON. CMC also shows high hit ratios in the EPAC. UM shows the lowest hit ratios in most regions with the exception of the WPAC.

During the DIMOSIC period, there were 16, 23, 34, and 27 TCs generated in the NATL, the EPAC, the WPAC, and the SHEM, respectively. However, not all observed TC geneses were predicted by the models. Figure 10a lists the number of TCs which were completely missed by models in each sub-region. It shows that JMA missed the genesis of total 30 TCs globally which is the most among all models. Most TC missed by JMA were in the WPAC and the SHEM. UM also missed many TC geneses in the EPAC. SHiELD had the least number of missed TCs globally followed by IFS and CMC. The newer IFS-47R3 missed more TCs than IFS in the NATL and the EPAC.

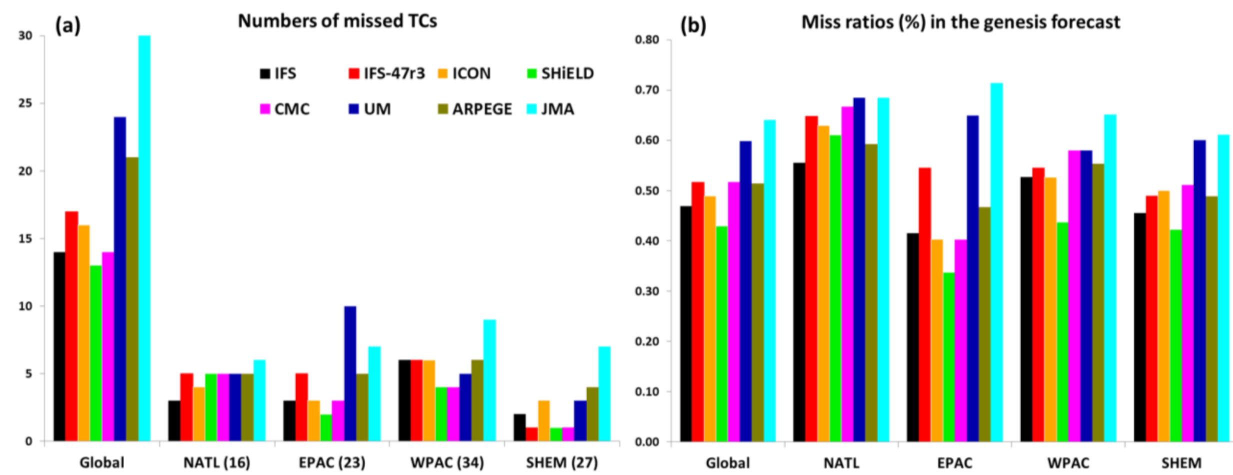

FIGURE 10. (a) Numbers of missed TCs and (b) miss ratios (%) in the genesis forecasts from the eight models. Abbreviations and colors used for the models and abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7. The numbers on the abscissa in (a) indicate the observed TC numbers in each sub-regions.

In the DIMOSIC period, models were initialized every three days. Hence, during the 10 days before an observed TC genesis event, a model could have three or four 10-day forecasts initialized and these runs are expected to predict this genesis event. Therefore, besides counting the number of completely missed TCs, the "miss ratio" can be computed as the number of missing cases compared to the number of expected genesis hit events (Chen et al. 2019b). The miss ratios for each model in the different sub-regions are shown in Fig. 10b to better revel the differences among the models. It shows that SHiELD shows the lowest miss ratios generally, except for in the NATL, while JMA and UM still struggle with relatively high miss ratios globally. We note that IFS-47R3 shows higher miss ratios than IFS in all sub-regions (Fig. 10b), including in the WPAC and the SHEM where IFS-47R3 shows the same or fewer numbers of completely missed TCs than IFS (Fig. 10a).

Following Chen et al. (2019b), beyond the scores of hit events, false alarms, and missing cases, we also investigate how precisely a model could predict the timing of TC genesis by comparing the "length of lead time" (see Fig. 8 in Chen et al. 2019b). The observed genesis lead time (OLT) is defined as the difference in time between the model initial time and the time at which observed TC genesis occurred. On the other hand, the time span from the model initial time to the model-predicted TC genesis lead time is referred as the model genesis lead time (MLT). The differences between the MLT and OLT (DMO) can indicate how accurate a model is in generating storms at the observed genesis time. If a model-predicted TC genesis occurred exactly at the observed TC genesis time, the DMO of this hit event is "zero". A positive DMO means that the model hit event occurs later than the observed TC genesis time, while negative DMO values are associated with early initiation of the TC in the model

For each observed TC, it is expected that more than one hit event will happen in the set of 10-day forecasts that cover the observed genesis time. To assess the predictive skill of each model, we only consider the maximum OLT, corresponding to the integration that identified the observed TC at the longest lead time. Figure 11a shows the mean values of the maximum OLT of all observed TCs in the four major sub-regions. In the NATL, both SHiELD and ARPEGE show a 110-hour OLT which is longer than the OLTs of other models, e.g. 78-hour OLT of UM. This indicates that SHiELD and ARPEGE could, on average, predict a hit TC genesis event 32 hours earlier than UM in the NATL. SHiELD also shows the earliest hit events in the EPAC and the WPAC, while ARPEGE shows the earliest hit events in the SHEM. The models in this study generally predict hit events earlier in the EPAC than in other sub-regions, except for JMA, which performs betters in the WPAC and the SHEM than in other sub-regions. It is also interesting to see that IFS shows earlier hit events than the newer IFS-47R3 in most sub-regions.

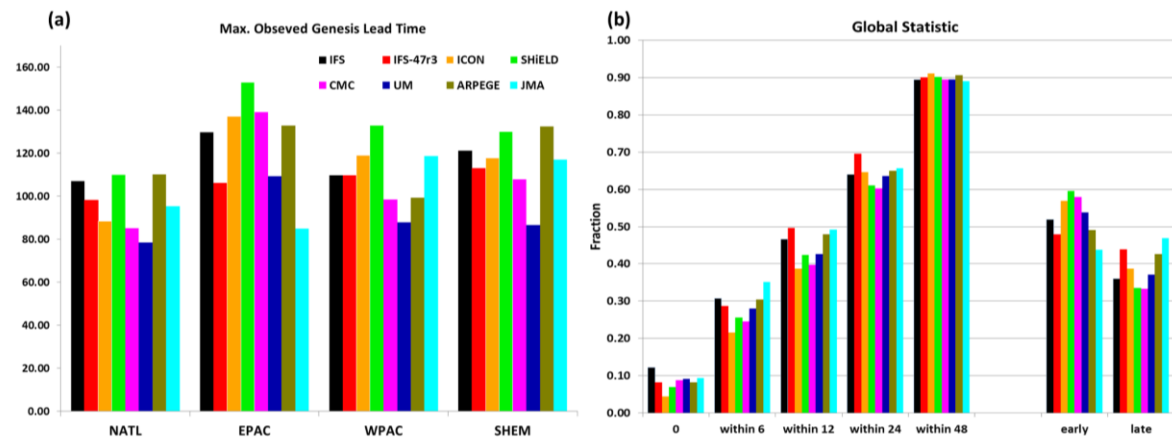

FIGURE 11. (a) Mean values of maximum observed genesis lead time (in hours) of all storms in each sub-region for the eight models. Abbreviations used for the sub-regions on the abscissa are the same as in Fig. 7. (b) Fractions of global total hit events in each model that occurred within a certain DMO length. On the abscissa, "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" is for all hit events with negative DMOs and "late" is for all hit evens with positive DMOs. Abbreviations and colors used for the models in both (a) and (b) are the same as in Fig. 7.

Figure 11b shows the fraction of global total hit events in each model which occurred within a certain length of DMO (indicated on the abscissa). From the definition of DMO, a model showing more genesis cases with short DMOs indicates that the model could predict more accurate genesis timings of its hit events. It can be found that IFS shows the highest fraction among all of the eight models in the "zero" DMO length category. The results indicate that IFS shows the highest ratio of its hit events forecasted at the observed TC genesis time among the models. Besides the two IFSs, JMA and ARPEGE also accurately predict TC genesis timings within the first three categories ("zero", "within 6", and "within 12") which is the 24-hour window (12 hours before or after) centered on the observed TC genesis time. In contrast, ICON shows the smallest fractions in the first three DMO length categories, which implies that the accuracy of TC genesis timing of ICON is relatively low compared to the other models.

From the result of the "within 48" DMO in Fig. 11b, we can see that more than 89% of hit events in each of the models occur within the 48 hours before or after the observed TC genesis time. When comparing the results in the "early" and "late" categories, it is seen that most models forecast their hit events before the observed TC genesis time, except for JMA. The ratios of "early" to "late" cases are larger in SHiELD and CMC compared to other models, while the IFS-47R3 shows relatively even number of cases of hit events generated before or after the observed TC genesis time.

4 Summary and Discussion

The DIMOSIC project provides a great opportunity to engage the worldwide community of medium-range modeling centers on cooperative model research and development. This study investigated TC forecast skills in the eight participating global medium-range forecast models during the year-long DIMOSIC period (June 2018 to June 2019). All models conducted 10-day forecasts from the same initial conditions based on the ECMWF IFS model cycle 45R1. The horizontal resolutions of the eight models ranged from 5 to 25km, and there were different choices of dynamical cores and physics parameterizations across the models (Magnusson et al. 2022). The forecast skills of TC track and intensity have been presented for the eight models. The TC-related precipitation and the performance of TC genesis forecasts have also been evaluated.

Comparing the model forecasts to the observations for the 109 TCs in the DIMOSIC period, IFS (45R1) and the updated version IFS-47R3 shows the best global averaged TC track forecasts, followed by ICON and SHiELD. CMC shows a relatively higher error than others before the 72-hour lead time. Based on our preliminary investigates, it could be related to the initializing moisture shock given that the CMC has much moister analyses than IFS which may induce convection collapses after the initialization with IFS ICs. For the TC track forecasts in different sub-regions, UM and ICON show the lowest track errors in the NATL, while SHiELD had the best track forecasts in the EPAC. In the WPAC and the SHEM, both IFS-47R3 and ICON show the lowest TC track errors, followed by SHiELD. From the analyses of along-track (AT) and cross-track (CT) errors, the models behave differently in different sub-regions. All models showed a slow-moving bias and a poleward bias in the NATL and the WPAC, but in the EPAC, except for JMA, most models show a fast-moving bias. In contrast, there are no consistently slow- or fast-moving biases among models in the SHEM.

For TC intensity forecasts, based on the TC tracker results using the interpolated 0.5 degree resolution data, SHiELD performs relatively better than other models, followed by UM and ARPEGE. From Table 2, we can see that the resolutions of the models range between 5 and 25 km, and the resolution of SHiELD is in the middle of that range. Therefore, the outperforming TC intensity by SHiELD may imply that the resolution is not the only major factor limiting TC intensity in global models. The use of dynamics and physics in the model also plays important role. The performance of TC track and intensity forecasts could reveal some of the characteristics of a model especially related to its dynamics and physics interactions. In Chen et al. (2019b), it has been demonstrated that updating the GFS dynamical core to the nonhydrostatic FV3 (Lin 2004; Putman and Lin 2007; Harris et al. 2020) can largely improve TC intensity forecasts, and additional improvements in TC intensity and genesis forecasts were seen when replacing the Zhao-Carr cloud microphysics scheme with the advanced GFDL cloud microphysics scheme (Zhou et al. 2019).

Here, we attempt to probe into the characteristics of the models based on their biases of TC track and intensity. Figure 12 shows the scatter plots of TC track and intensity errors for all of the forecasts during the 72-120-hour lead time in each model. Some similarities can be found in the scattered distributions of the two IFSs and ICON, including both ranges of intensity bias and track error. This is consistent with the findings in Magnusson et al. (2022) that IFS and ICON behave relatively similarly due to the sharing of partial physical parameterizations sharing between ECWMF and DWD. In contrast, SHiELD, ARPEGE, and JMA show rather unique patterns their own.

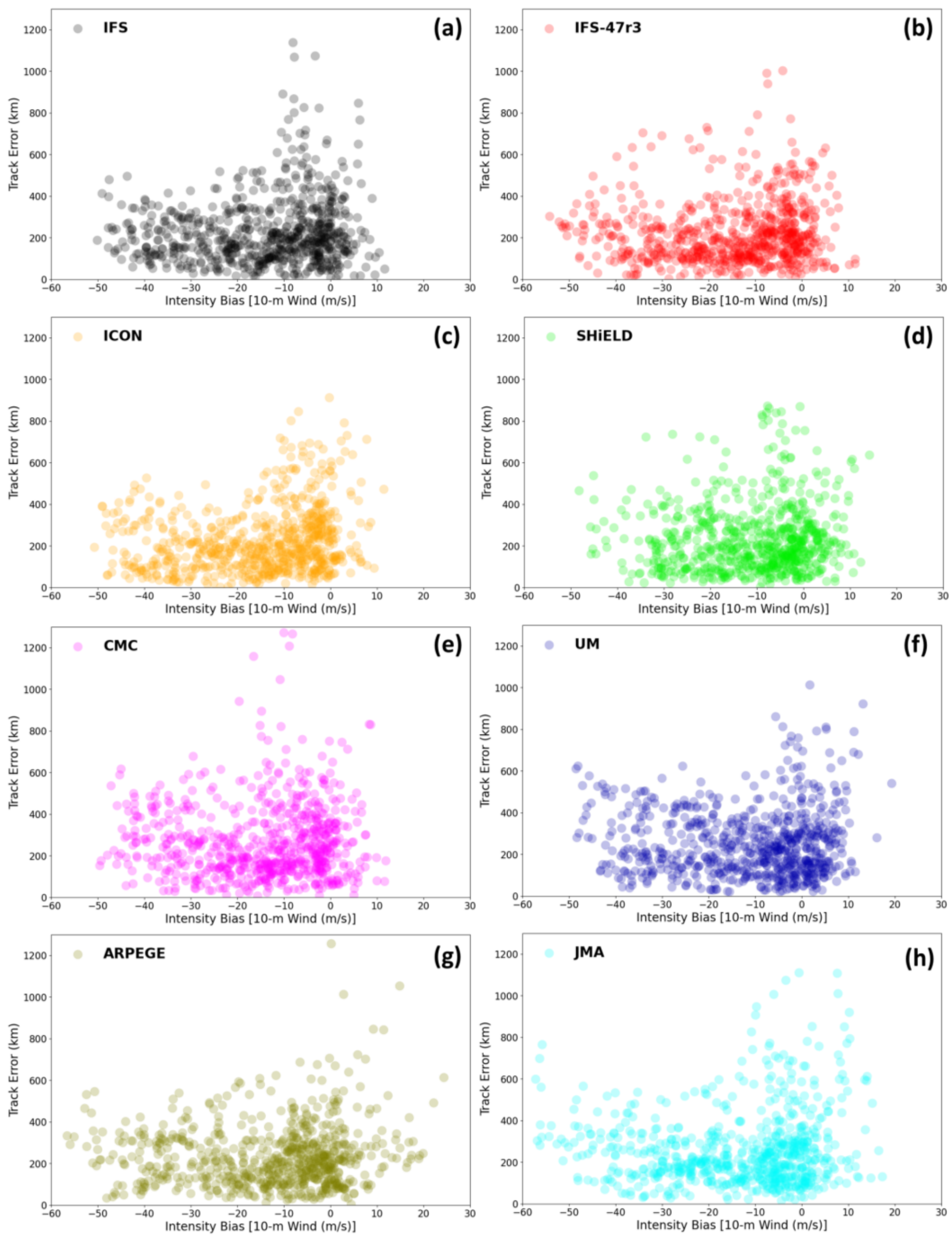

FIGURE 12. Scatter plot distribution of track errors (unit: km; ordinate) and intensity biases (the maximum 10-m wind speed; unit: m s^{-1}; abscissa) of all forecasts during the lead time of 72-120 hour for (a) IFS, (b) IFS-47r3, (c) ICON, (d) SHiELD, (e) CMC, (f) UM, (g) ARPEGE, and (h) JMA.

Figure 12 also shows that in most models, forecasts with larger track error (>700 km) are usually accompanied by smaller intensity biases (<10 ms^{-1}). In contrast, for those forecasts with larger intensity biases, their track errors are not consistently larger. At GFDL, it has been noticed that when the performance of TC track forecasts was improved by using an advection scheme with a stronger damping in the dynamics, a degradation of TC intensity was observed. 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 (Gao et al. 2021). This was attributed to the impact of stronger damping, which suppresses finer-scale features and activities, e.g. grid-scale convection in the TCs, which further suppresses the TC intensities. An in-depth study of the impact of grid-scale convection activity on TC track forecasts in SHiELD is in preparation.

Since the interpolated data cannot fully represent the performance of TC intensity forecasts in the models, the TC-related precipitation was also evaluated to provide another perspective on forecasted TC characteristics. Compared to the GPM observation data, all models under-predict the amount of TC-related precipitation, especially in the Pacific Ocean. UM better captures the regions with annual accumulated TC-related precipitation of more than 300mm compared to the other models. However, when considering all TC-related precipitation areas, the ETS of UM is generally lower than that of other models. This could be related to the relatively large track errors of UM, especially in the EPAC. In contrast, SHiELD shows the largest dry bias in TC-related precipitation among all models, but it still achieves relatively high ETSs in the EPAC due to its better track forecasts in this sub-region. As to the intensity of TC-related precipitation, most models over-predict medium precipitation events but under-predict light and heavy precipitation events. Among all models, SHiELD noticeably over-predicts precipitation events at the intensity of 1-5 mm $(6\ h)^{-1}$. The SHiELD development team at GFDL will take a close look at its low precipitation amount and over-predicted medium intensity precipitation events in the future.

The assessment of TC genesis forecast skill was based here on hits and misses, measures that showed significant inter-model variability across the different sub-regions of interest. All models show the highest hit ratios with the fewest total genesis events in the EPAC, which indicates that models can better predict TC genesis in the EPAC. In contrast, models generate more false alarms in the SHEM. SHiELD shows the highest hit ratios globally, followed by the two IFSs and ICON. CMC also achieves high hit ratios in the EPAC. In contrast, UM shows lower hit ratios than other models. As for the missed TC genesis cases, JMA missed 30 of the 100 observed TC geneses during the target period, which is the most among all models. Note that JMA uses the coarsest model resolution among all participating models, which could impair its TC genesis performance. In contrast, SHiELD shows the least number of missed TCs globally, which may be benefiting from its better TC intensity forecast.

We also investigate how well the participating models predict the timing of TC genesis by comparing the “length of lead time” proposed by Chen et al. (2019b). The results show that models can generally accurately predict TC formation earlier in the EPAC than in other sub-regions, except for JMA which predicts the WPAC and the SHEM hit events earlier than in other sub-regions. SHiELD generally predicts the earliest hit events globally, while ARPEGE also predicts TC genesis events earlier than other models in the NATL and the SHEM. Based on the differences between the model genesis lead time and the observed genesis lead time, IFS shows the most accurate timing of the TC genesis forecast, followed by JMA and ARPEGE. In contrast, the accuracy of genesis timing in ICON is relatively lower than that of other models. We also found that most models develop TCs earlier than observed in the best track, except for JMA. In addition, more than 89% of hit events in each of the models occur within 48 hours of the observed genesis time.

The comparison between IFS (version 45R1) and IFS-47R3 provides an opportunity to examine the incremental change obtained for an upgrade of one model. The upgrade from version 45R1 to 47R3 includes many changes in data assimilation and model physics. The changes and meteorological impacts have been documented by ECMWF on the website: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model. One of the listed impacts from the upgrade is the improvement of TC position errors. From our analysis, the average track errors during the first 120 hours of IFS-47R3 are 2.5 to 10.8 km less than IFS (Fig. 3) in the four major sub-regions, which is consistent with the ECMWF implementation report. However, from the analyses of the along-track and cross-track errors, the biases of slow and poleward movement are similar in these two model versions. Possibly associated with the major upgrade to moist physics (Bechtold et al. 2020), IFS-47R3 shows a slightly larger negative TC intensity biases than the older IFS version which was also found in Magnusson et al. (2021). Although total precipitation predictions remain similar across the two IFS versions, the newer IFS-47R3 achieves higher ETS scores for large accumulations, especially in the WPAC and the SHEM. We also found that IFS-47R3 has more precipitation events with stronger precipitation intensity (10-50 mm $(6\ h)^{-1}$) than IFS. However, as to the TC genesis forecast, IFS-47R3 shows some degradation from IFS, including more missed TC genesis event, higher miss ratios, shorter genesis lead times, and less accuracy on TC genesis timing. These degradations in TC genesis performance may be related to the weaker TC intensities in the newer version.

To summarize, in this study, extensive evaluation was made of the performance of TC forecasts in the DIMOSIC models based on one year of model predictions initialized with the same initial conditions. Although it is hard to precisely isolate the influence of individual components in different model formulations on TC forecast skill in such an overview, the comparisons based on different evaluation metrics highlight important similarities and differences between the models. The results will be valuable for model developers in participating centers as a benchmark of TC forecast skill with the impact of the initial condition quality removed. Also, common forecast biases of the TC movement and TC-related precipitations indicate general deficiencies in DIMOSIC models and point out a direction for model developers for further model improvement.

Acknowledgments

The authors thank Kun Gao, Jie Chen, Morris Bender, and Tom Knutson for GFDL internal review, and Lucas Harris, James Doyle, and Simon Lang for their comments helped to improve this article. Authors also would like to thank other DIMOSIC participants, Duncan Ackerley, Yves Bouteloup, K. C. Kwon, Yoonjin Lim, Mio Mastueda, Takumi Matsunobu, and Yamaguchi Munehiko for their contribution to the DIMOSIC project.

Open Research

All DIMOSIC model interpolated data can be requested from Linus Magnusson. All TC analyses are archived in the GFDL Tape Archive System at /archive/jhc/DIMOSIC/Analysis/TC and can be requested from Jan-Huey Chen.

References

Becthold, P., R. Forbes, I. Sandu, S. Lang, & M. Ahlgrimm (2020), A major moist physics upgrade for the IFS. *ECMWF Newsletter, 164*, URL: https://www.ecmwf.int/en/newsletter/164/meteorology/major-moist-physics-upgrade-ifs

Chen, J.-H., S.-J. Lin, L. Magnusson, M. A. Bender, X. Chen, L. Zhou, B. Xiang, S. L. Rees, M. J. Morin, & L. Harris (2019a), Advancements in hurricane prediction with NOAA's next generation forecast system. *Geophysical Research Letters, 46* (8), 44954501, doi:10.1029/2019GL082410.

Chen, J.-H., Lin, S., Zhou, L., Chen, X., Rees, S., Bender, M., & Morin, M. (2019b), Evaluation of Tropical Cyclone Forecasts in the Next Generation Global Prediction System, *Monthly Weather Review, 147*(9), 3409-3428, doi: 10.1175/MWR-D-18-0227.1

DWD (2022), ICON : Icosahedral Nonhydrostatic Weather and Climate Model. *Technical Report*, *DWD*. URL: https://code.mpimet.mpg.de/projects/iconpublic/wiki/Documentation

ECMWF (2018), IFS Documentation CY45R1. *Technical Report*, *ECMWF*. URL: https://www.ecmwf.int/en/publications/ifs-documentation

ECMWF (2021), IFS Documentation CY47R3. *Technical Report*, *ECMWF*. URL: https://www.ecmwf.int/en/publications/ifs-documentation

Gao, K., L. Harris, L. Zhou, M. A Bender, and M. Morin (2021), On the sensitivity of hurricane intensity and structure to horizontal tracer advection schemes in FV3. *Journal of the Atmospheric Sciences*, 78(9), doi:10.1175/JAS-D-20-0331.13007-3021.

Girard, C., & Coauthors (2014), Staggered vertical discretization of the Canadian Environmental Multiscale (GEM) model using a coordinate of the log-hydrostatic-pressure type. *Monthly Weather Review, 142* (3), 1183.

Harris, L. M., S.-J. Lin, & C. Tu (2016), High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *Journal of Climate, 29*, 4293–4314, doi:10.1175/JCLI-D-15-0389.11196, doi:10.1175/MWR-D-13-00255.1.

Harris, L., & Coauthors (2020), GFDL SHiELD: A unified system for weather to seasonal prediction. *Journal of Advances in Modeling Earth Systems*, *12* (10), doi:10.1029/2020MS002223.

Hong, Y., K.-L. Hsu, S. Sorooshian, & X. Gao (2004), Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *Journal of Applied Meteorology*, *43* (12), 18341853, doi:10.1175/JAM2173.1.

JMA (2019), Outline of the operational numerical weather prediction at the Japan Meteorological Agency, *Technical Report, Japan Meteorological Agency*. URL http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2019-nwp/index.htm.

Lin, S.-J. (2004), A “vertically Lagrangian” finite-volume dynamical core for global models, *Monthly Weather Review*, 132, 2293-2307.

Magnusson, L., J.-H. Chen, S.-J. Lin, L. Zhou, & X. Chen (2019), Dependence on initial conditions vs. model formulations for medium-range forecast error variations. *Quarterly Journal of the Royal Meteorological Society*, *145*, doi:10.1002/qj.3545

Magnusson, L., & co-authors (2021), Tropical cyclone activities at ECMWF. *ECMWF Technical Memorandum 888*

Magnusson, L., D. Ackerley, Y. Bouteloup, J.-H. Chen, J. Doyle, P. Earnshaw, Y. C. Kwon, M. Köhler, S. T. K Lang, Y.-J. Lim, M. Matsueda, T. Matsunobu, R. McTaggart-Cowan, A. Reinecke, M. Yamaguchi1, & L. Zhou (2022), Skill of medium-range forecast models using the same initial conditions. *Bulletin of the American Meteorological Society*, *103*(9), E2050-E2068, doi:10.1175/BAMS-D-21-0234.1

McTaggartCowan, R., & Coauthors (2019), Modernization of atmospheric physics parameterization in Canadian NWP. *Journal of Advances in Modeling Earth Systems, 11* (11), 35933635, doi:10.1029/2019MS001781.

Miller, R. J., A. J. Scrader, C. R. Sampson, & T. L. Tsui (1990), The Automated Tropical Cyclone Forecast System (ATCF). *Weather and Forecasting*, *5*, 653–660.

Mogensen, K. S. , L. Magnusson , & J.-R. Bidlot (2017), Tropical cyclone sensitivity to ocean coupling in the ECMWF coupled model: Tropical cyclone sensitivity. *J. Geophys. Res. Oceans*, *122*, 4392–4412, doi:10.1002/2017JC012753.

Pollard, R. T. , P. B. Rhines , & R. O. R. Y. Thompson (1973), The deepening of the wind-mixed layer. *Geophys. Fluid Dyn.* , *4*, 381–404, doi:10.1080/03091927208236105.

Putman, W. M., & S.-J. Lin (2007) Finite-volume transport on various cubed-sphere grids. *Journal of computational physics*, 227, 55-78, doi:10.1016/j.jcp.2007.07.022.

Rabier, F., H. Ja¨rvinen, E. Klinker, J.-F. Mahfouf, & A. Simmons (2000), The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society, 126* (564), 24 1143–1170, doi:10.1002/qj.49712656415

Roehrig, R., & Coauthors (2020), The cnrm global atmosphere model arpegeclimat 6.3: Description and evaluation. *Journal of Advances in Modeling Earth Systems*, *12* (7), doi:10.1029/2020MS002075.

Schaefer, J. T. (1990) The critical success index as an indicator of warning skill. *Weather and Forecasting, 5* , 570–575.

Sampson, C. R., & A. J. Schrader (2000), The Automated Tropical Cyclone Forecasting System (version 3.2). *Bulletin of the American Meteorological Society*, *81*, 1231–1240.

Walters, D., & Coauthors (2019), The met office unified model global atmosphere 7.0/7.1 and jules global land 7.0 configurations. *Geoscientific Model Development*, *12*(5), 19091963, doi:10.5194/gmd-12-1909-2019.

Zeng, X. , & A. Beljaars (2005), A prognostic scheme of sea surface skin temperature for modeling and data assimilation: Sea surface skin temperature scheme. *Geophysical Research Letters, 32*, L14605, doi:10.1029/2005GL023030.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, & S. Rees (2019), Toward convective scale prediction within the Next Generation Global Prediction System. *Bulletin of the American Meteorological Society*. 100 (7): 1225-43, doi:10.1175/BAMS-D-17-0246.1

Journal of Advances in Modeling Earth Systems Supporting Information for

Tropical Cyclone Forecasts in the DIMOSIC Project

Jan-Huey Chen[1,2], Linjiong Zhou[3], Linus Magnusson[4], Ron McTaggart-Cowan[5], and Martin Köhler[6]

[1]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

[2]University Corporation for Atmospheric Research, Boulder, CO, USA

[3]Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, NJ, USA

[4]European Centre for Medium-Range Weather Forecasts, Reading, UK

[5]Environment and Climate Change Canada, Montreal, Canada

[6]Deutsche Wetterdienst, Offenbach, Germany

Contents of this file

Introduction

This supporting information file includes:

1) Analyses of TC track errors of GFDL SHiELD in the North Atlantic basin (NATL) (Text S1 and Figures S1-S3)

2) Analyses of along-track (AT) and cross-track (CT) errors for all 8 models in the 4 major sub-regions (Text S2 and Figures S4-S7)

Text S1.

During the DIMOSIC period, SHiELD is one of the top performance models which show low tropical cyclone (TC) track forecast errors in most sub-regions but not in the North Atlantic basin (NATL). We therefore further investigate the large track error of SHiELD in the NATL to supplement the results of track error analysis, and to provide references for the SHiELD development team at GFDL.

Figure S1a shows the mean TC track forecast errors for all of the 8 DIMISIC models in the NATL, with the differences of TC track errors of the 7 models comparing to IFS (Fig. S1b). It can be found that SHiELD shows a much larger track error than most other models during the 72-120 lead times.

We investigated the track forecasts of the 16 TCs in the NATL during the target year individually and found that the basin-wide mean track errors are actually dominated by 2 TCs, Hurricanes Florence (2018) and Leslie (2018). Comparing to other leading models, SHiELD shows difficulty to forecast the storm movements of these two hurricanes. Figure S2 shows the forecasted tracks of Florence from IFS-47R3, SHiELD, ICON, and UM initialized at 00Z Sep. 1st (Fig. S2a) and 00Z Sep. 13th (Fig. S2b) comparing to the best track. It can be found that SHiELD shows slow moving biases on Florence's translation in both early and late stage of its lifetime, while IFS-47R3, ICON and UM do not show similar forecast biases. Different from Florence which had a steady westward movement crossing the Atlantic Ocean, Leslie had an irregular track with many loops in the middle of Atlantic Ocean before heading northeasterly to Portugal. From Fig. S3, we can see that SHiELD didn't well capture some of the sharp turns which are critical in the track forecast of this storm. In contrast, IFS-47R3, ICON, and UM can perform much better "v" shape turns in their forecasts initialized from 00Z 1stOct. and from 00Z 7th Oct.

To summarize, the large TC track error of SHiELD in the NATL results from the slow moving bias shown in the forecasts of Hurricane Florence and the bias of direction of motion shown in the forecasts of Hurricane Leslie.

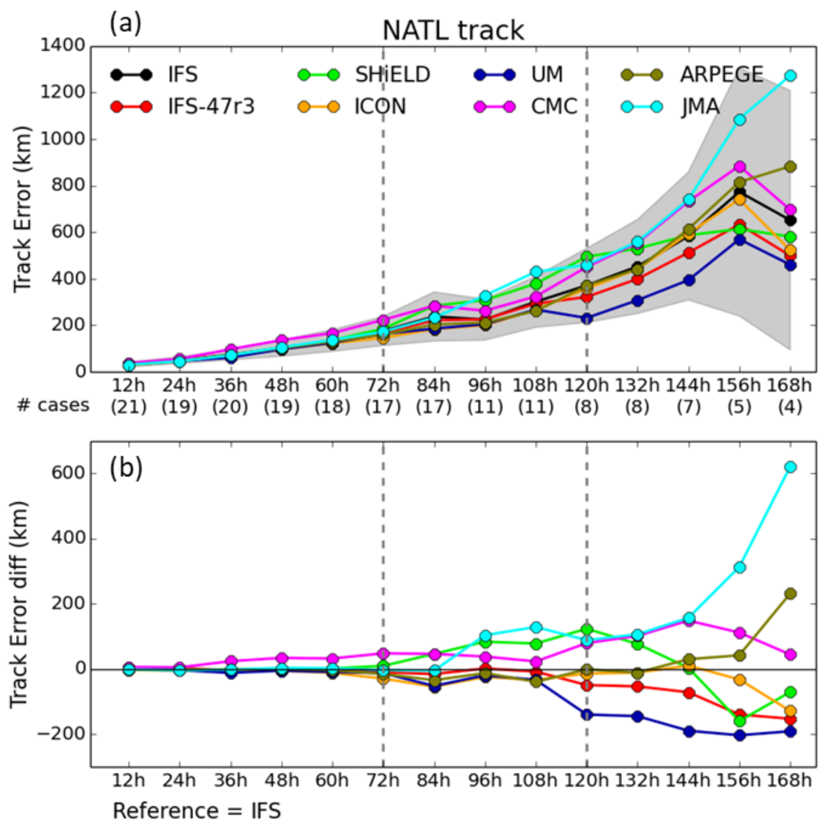

Figure S1. (a) Mean TC track forecast errors (km) along with the model forecast lead time for IFS (black), IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), CMC (magenta), APREGE (grass green), and JMA (light blue) in the North Atlantic basic (NATL). The 95% confidence levels for IFS are indicated by the gray 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. (b) Mean TC track forecast error differences of IFS-47r3 (red), SHiELD (green), ICON (yellow), UM (blue), CMC (magenta), APREGE (grass green), and JMA (light blue) in the NATL comparing to IFS.

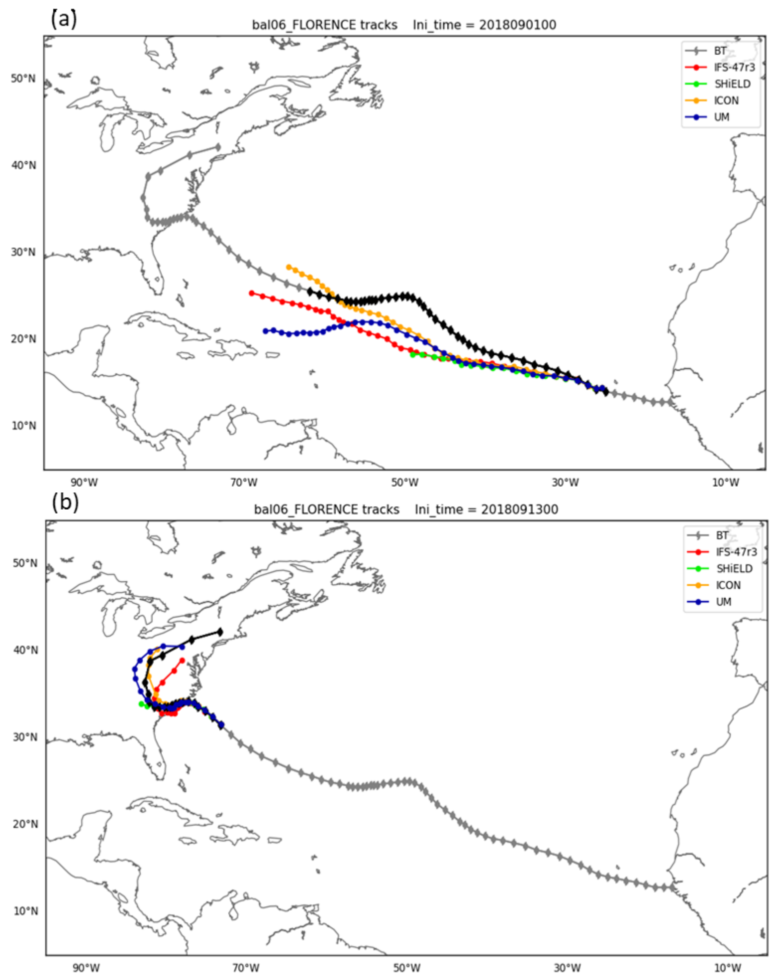

Figure S2. (a) Model forecasted tracks from IFS-47r3 (red), SHiELD (green), ICON (yellow), and UM (blue) initialized at 00Z 20180901 comparing to the Automated Tropical Cyclone Forecast (ATCF) best track for Hurricane Florence. Dots for model forecasts and typhoon symbols for the best track are at 6-hour interval. The best track of Hurricane Florence during the 10-day forecast period is plotted in black, and the rest part of the best track is plotted in grey. (b) As in (a), but for model forecasts initialized at 00Z 20180913.

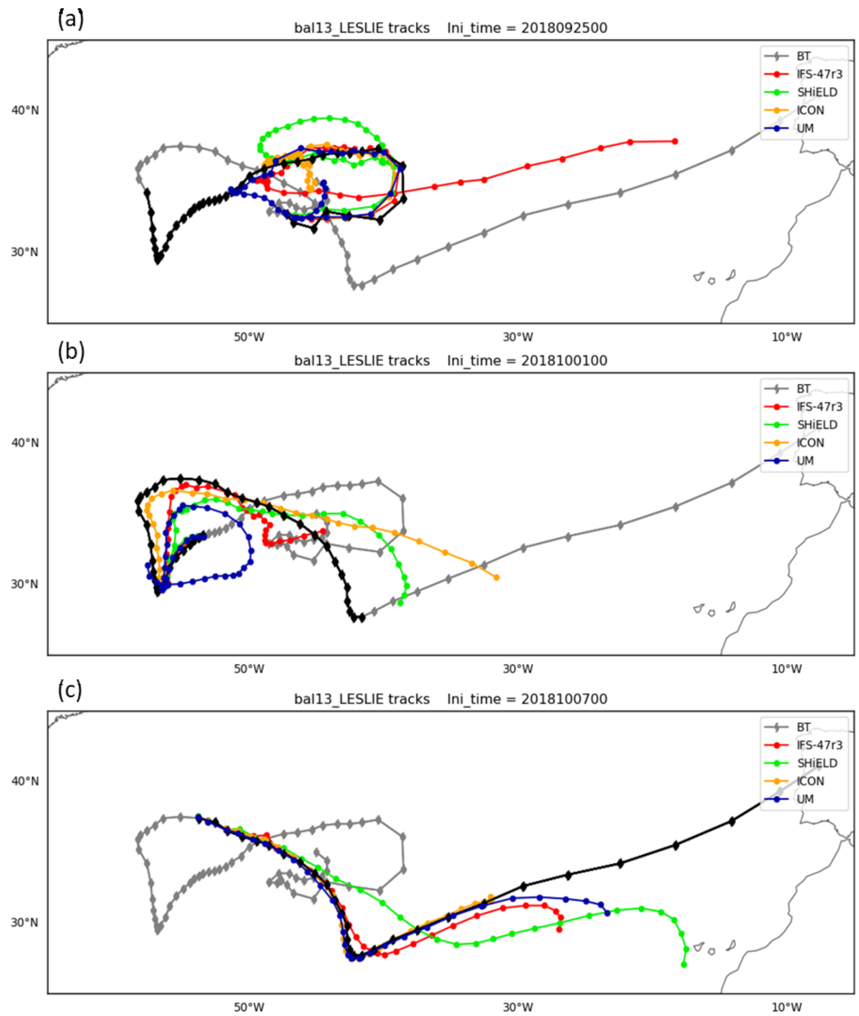

Figure S3. As in Figure S2, but for the model forecasts initialized at (a) 00Z 20180925, (b) 00Z 20181001, and (c) 00Z 20181007 for Hurricane Leslie comparing to the best track.

Text S2.

Analyses of along-track (AT) and cross-track (CT) errors for all eight models in the four major sub-regions are shown in Figs. S4-S7. We note that the models behave differently in different sub-regions. In the NATL (Fig. S4), the AT and CT errors during the first 5 days evenly contribute to the total track errors in the two IFSs. UM AT/CT skill is similar to that of the IFSs but with a larger AT error component on Day 3 to 4 (Fig. S4f). In contrast, ICON and CMC have a much larger AT error component than the CT error (Figs. S4c,e), which indicates that their total track errors in the NATL are mostly due to slow moving biases. The source of track error of SHiELD in the NATL is a combination of AT error during 96 to 132-hour lead time and CT error at Days 6 to 7 (Fig. S4d), which is consistent with the single case analyses (Text S1 and Figs. S1-S3).

In distinct behavior from the NATL, seven of the eight models show positive AT error biases in the EPAC during the first five days (Fig. S5). This indicates that the model-predicted TCs are usually moving too fast in this basin. In contrast, JAM suffers a slow-moving bias as shown by a dominant negative AT error bias (Fig. S5h). Also, the large track error of UM during the first five days in this region (Fig. 3) is mostly from a fast-moving bias, but a southward-moving bias also contributes. In the WPAC, all models consistently show negative AT error biases and positive CT error biases after Day 5 (Fig. S6). It is clear that a slow-moving bias is the main source of the total track errors in the two IFSs in this region (Figs. S6a,b). UM and JMA show large contributions of AT errors during Days 4-5, but northward-moving biases also contribute to the total error substantially (Figs. S6f,h). ICON, SHiELD, and CMC show even greater contributions of AT and CT errors to their total track errors than other models (Figs. S6c-e). In the SHEM (Fig. S7), except for SHiELD showing a relatively even contribution of AT and CT errors with small biases, other models generally exhibit much larger components of AT errors than CT errors. However, there are no consistently slow- or fast-moving biases among models in this basin. IFSs and CMC show consistently slow-moving biases after three days, while UM shows fast-moving biases in the later lead times. ICON, APREGE, and JMA show mixed slow- or fast-moving biases in the seven days of lead time.

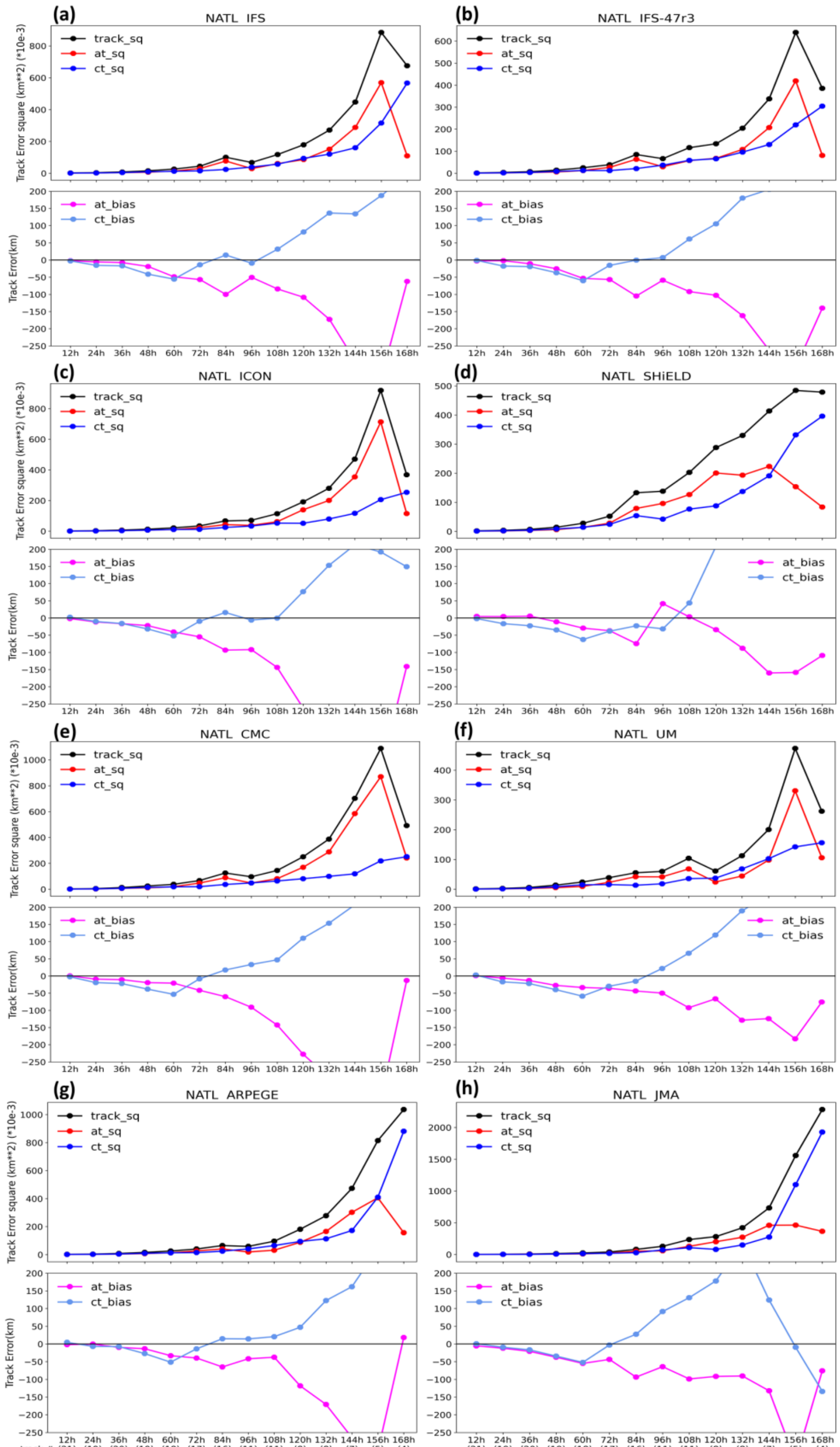

Figure S4. Analyses of along-track (AT) error and cross-track (CT) error for (a) IFS, (b) IFS-47R3, (c) ICON, (d) SHiELD, (e) CMC, (f) UM, (g) ARPEGE, and (h) JMA in the North Atlantic basin (NATL). The squares of total track errors (black), along-track errors (red), and cross-track errors (blue) are in the upper panels for each model. The biases of along-track (magenta) and cross-track (light blue) errors are in the lower panels. Numbers of homogeneous cases for each lead time are listed at the bottom of lower panels.

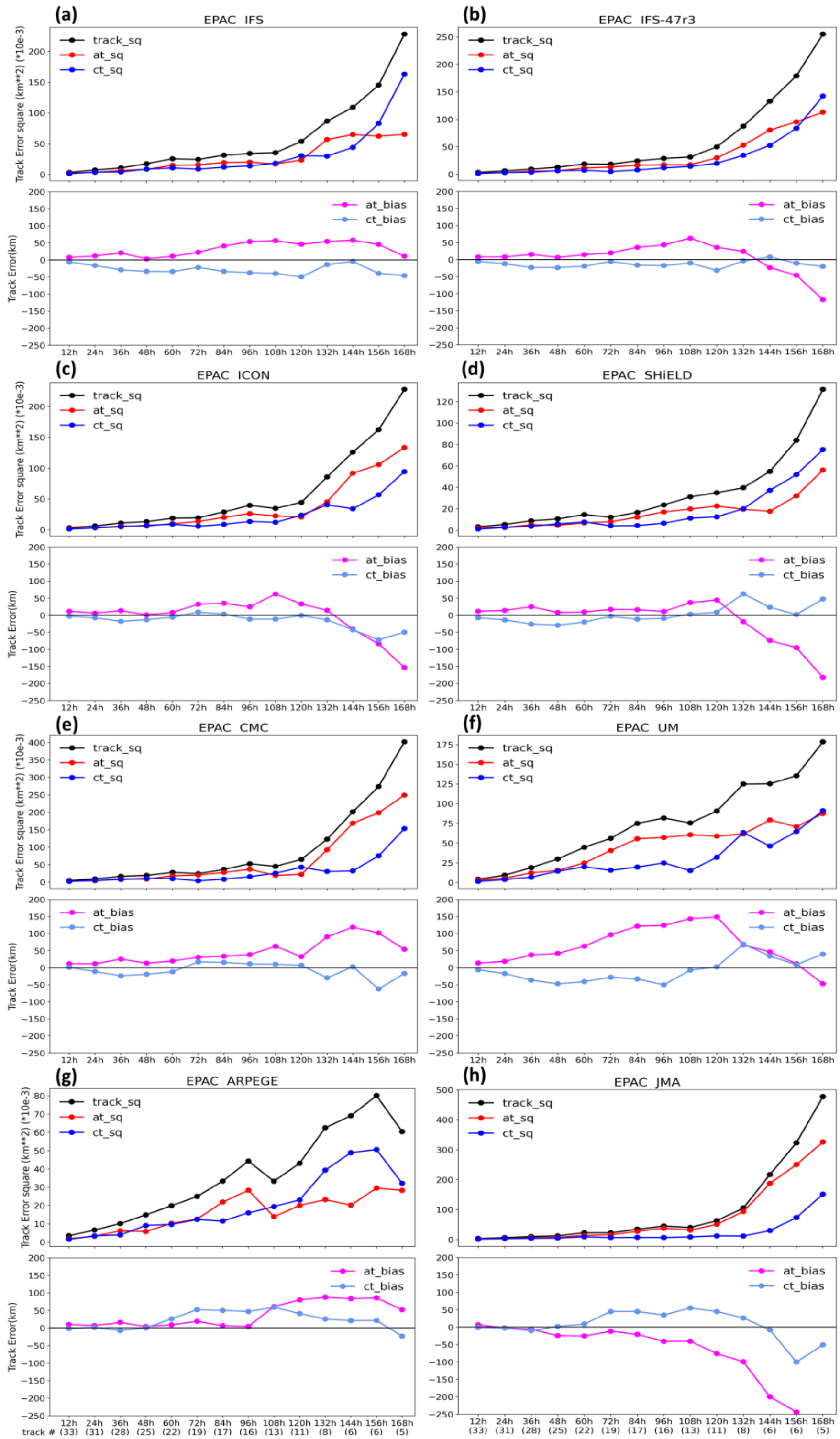

Figure S5. As in Figure S4, but for the analyses in the northeast Pacific basin (EPAC).

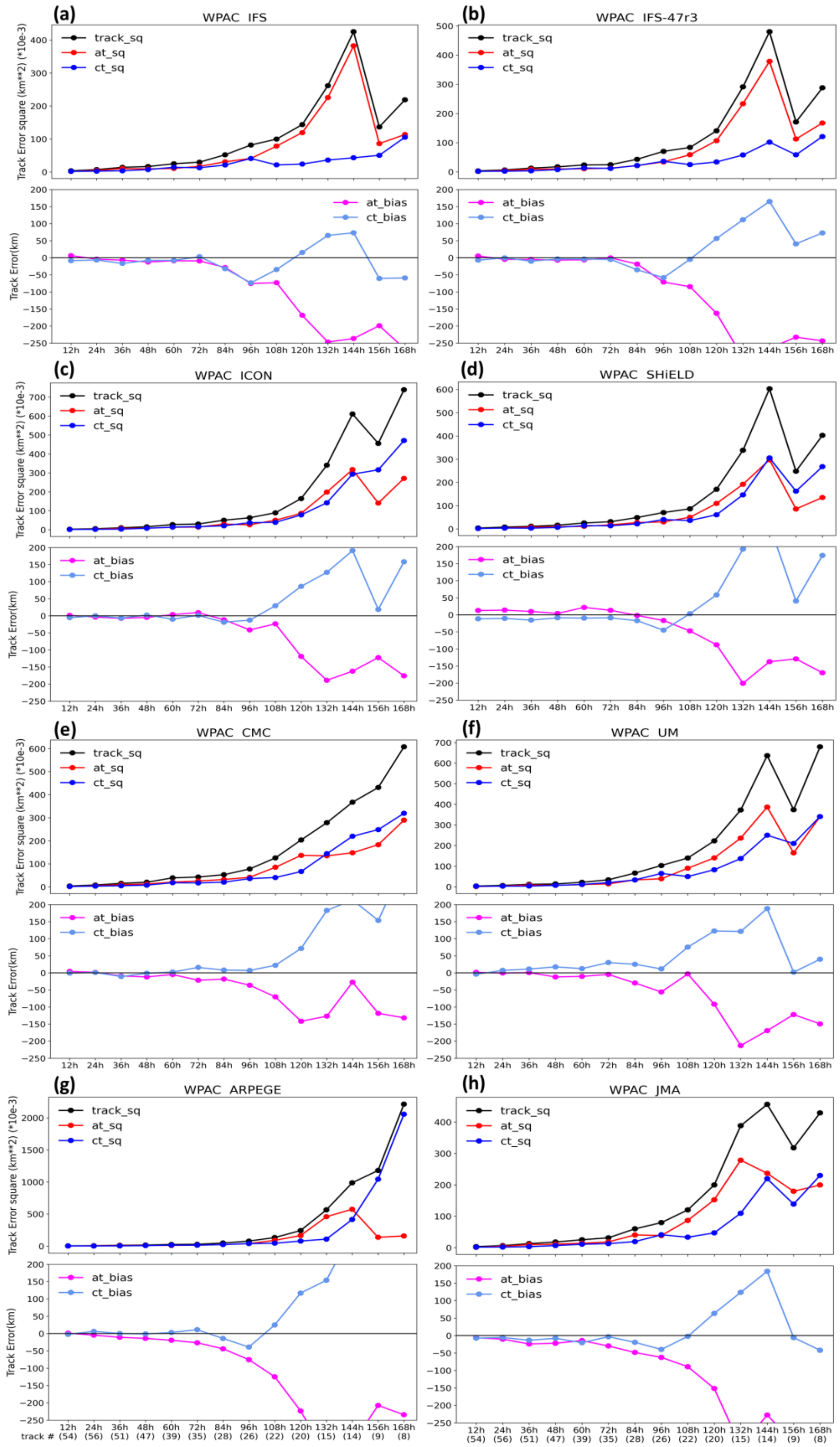

Figure S6. As in Figure S4, but for the analyses in the northwest Pacific basin (WPAC).

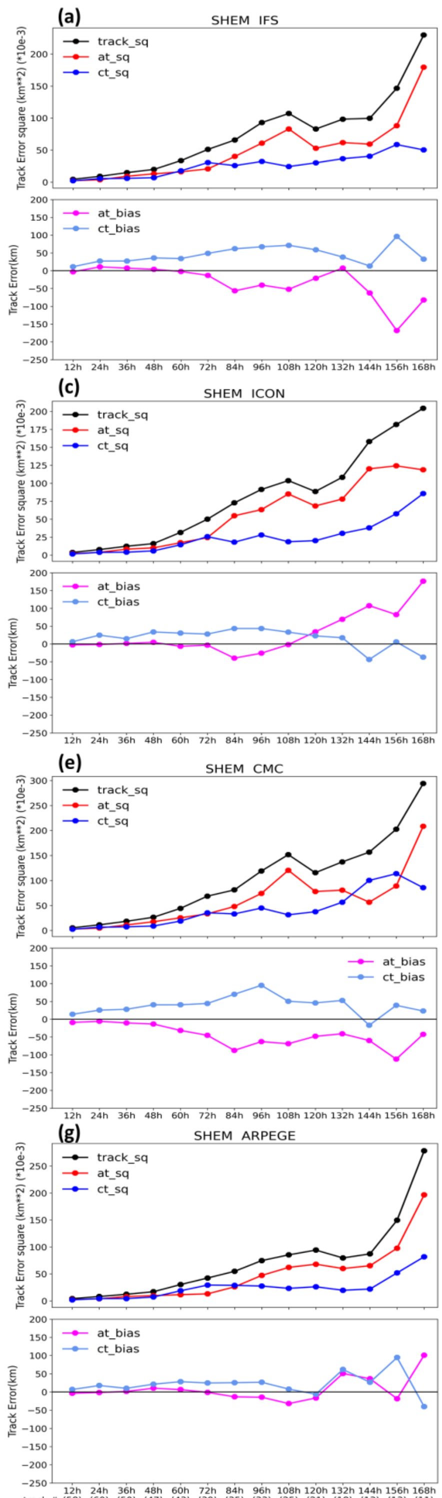

Figure S7. As in Figure S4, but for the analyses in the Southern Hemisphere (SHEM).