Impacts of Assimilating Additional Reconnaissance Data on Operational GFS Tropical Cyclone Forecasts

JASON A. SIPPEL,[a] XINGREN WU,[b,d] SARAH D. DITCHEK,[a,c] VIJAY TALLAPRAGADA,[d] AND DARYL T. KLEIST[d]

[a] *NOAA/Atlantic Oceanographic and Meteorological Laboratory/Hurricane Research Division, Miami, Florida*
[b] *I.M. Systems Group, Rockville, Maryland*
[c] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[d] *NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 29 March 2022, in final form 1 June 2022)

ABSTRACT: This study reviews the recent addition of dropwindsonde wind data near the tropical cyclone (TC) center as well as the first-time addition of high-density, flight-level reconnaissance observations (HDOBs) into the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). The main finding is that the additional data have profound positive impacts on subsequent TC track forecasts. For TCs in the North Atlantic (NATL) basin, statistically significant improvements in track extend through 4–5 days during reconnaissance periods. Further assessment suggests that greater improvements might also be expected at days 6–7. This study also explores the importance of comprehensively assessing data impact. For example, model or data assimilation changes can affect the so-called "early" and "late" versions of the forecast very differently. It is also important to explore different ways to describe the error statistics. In several instances the impacts of the additional data strongly differ depending on whether one examines the mean or median errors. The results demonstrate the tremendous potential for further improving TC forecasts. The data added here were already operationally transmitted and assimilated by other systems at NCEP, and many further improvements likely await with improved use of these and other reconnaissance observations. This demonstrates the need of not only investing in data assimilation improvements, but also enhancements to observational systems in order to reach next-generation hurricane forecasting goals.

SIGNIFICANCE STATEMENT: This study demonstrates that data gathered from reconnaissance missions into tropical cyclones substantially improves tropical cyclone track forecasts.

KEYWORDS: Aircraft observations; Dropsondes; Numerical weather prediction/forecasting; Data assimilation; Forecast verification/skill; Hurricanes/typhoons; Tropical cyclones

1. Introduction

With a goal of improving tropical cyclone (TC) forecasts, the National Centers for Environmental Prediction (NCEP) and the former National Meteorological Center (NMC) have assimilated airborne reconnaissance data into various operational models for about four decades. Over the years, models have gradually assimilated more data to the point that some, such as NCEP's Hurricane Weather Research and Forecasting (HWRF) model, now use most reconnaissance data transmitted in real time. Until recently, however, the NCEP Global Forecast System (GFS) used only a limited amount of available data. This paper will review the recent addition of dropwindsonde wind data near the TC center as well as the first-time addition of high-density, flight-level reconnaissance observations (HDOBs) into GFS.

Among reconnaissance data types, dropwindsondes have the longest history of operational assimilation. In the early 1980s (e.g., Burpee et al. 1984), airborne missions transmitted a limited amount of dropwindsonde data for operational use. The volume of dropwindsonde data assimilated had increased by the early–mid-1990s, and Burpee et al. (1996) found that they improved TC track forecasts up to 30%, which was about as big as the entire improvement in track forecasts from 1970 to 1990. These impressive results led NOAA to invest more heavily in TC reconnaissance, including the purchase of a Gulfstream-IV (G-IV) jet for synoptic surveillance missions. The G-IV missions also improved operational track forecasts, though by somewhat smaller amounts (up to 15% in Aberson 2010). Ditchek et al. (2021, manuscript submitted to *Wea. Forecasting*) reviewed the history of dropwindsonde impact assessments from 1992 to 2019 and estimated a median track improvement across various research and operational modeling systems of about 7%–8%.

A dedicated effort to assimilate other types of reconnaissance data, including high-resolution inner-core observations, began around 2008. A number of studies with research-quality regional data assimilation (DA) systems (Zhang et al. 2011; Weng and Zhang 2012; Aberson et al. 2015; Weng and Zhang 2016, hereafter WZ16) began to show considerable improvements in TC track and intensity forecasts associated with assimilating airborne Doppler radar velocity data as well as flight-level HDOBs (e.g., temperature, wind, and humidity) and observations from the stepped frequency microwave radiometer (SFMR). The most extensive of these studies, WZ16, showed that assimilating dropwindsondes and HDOBs together improved track and intensity forecasts by 10%–15%. The operational HWRF has also recently advanced in both model physics and DA, enabling use of all operationally transmitted dropwindsonde data, HDOBs, and airborne

Corresponding author: Jason A. Sippel, jason.sippel@noaa.gov

DOI: 10.1175/WAF-D-22-0058.1

TABLE 1. Specified observation errors used in GFSv16 for HDOB and dropwindsonde (DROPS) data for each given layer for winds (UV), temperature (T), and relative humidity (RH). In layers where the assumed error does not vary, only a single value is given. If the assumed error varies through a layer, then the error is expressed as a range from the bottom to the top of the layer. For moisture, error is expressed in terms of relative humidity but changed to specific humidity within GSI.

Variable	Surface–900 hPa	900–700 hPa	700–500 hPa	500–300 hPa	300–100 hPa
DROPS UV	2.4 m s^{-1}	2.4 m s^{-1}	2.4–2.8 m s^{-1}	2.8–3.4 m s^{-1}	3.4–2.5 m s^{-1}
DROPS T	1.2–0.9 K	0.9–0.8 K	0.8 K	0.8–0.9 K	0.9–1.2 K
DROPS RH	20%	20%	20%	20%	20%
HDOB UV	5.5 m s^{-1}	5.5 m s^{-1}	5.5 m s^{-1}	5.5 m s^{-1}	5.5 m s^{-1}
HDOB T	2.5–2.1 K	2.1–1.7 K	1.7 K	1.7 K	1.7 K
HDOB RH	20%	20%	20%	20%	20%

Doppler radar data. NCEP last assessed the reconnaissance impact for HWRF in a 2019 version of the model and found that it improved the intensity forecast by 10%–15% through 72-h lead time (Zawislak et al. 2021, their Fig. 4).

The types of reconnaissance data assimilated into the NCEP GFS, however, were still somewhat limited through 2020. Only dropwindsonde data were assimilated, and dropwindsonde wind data close to the center of any TC was discarded due to concerns of forecast degradation since dropwindsonde drift (e.g., lateral advection) is not accounted for in GFS (e.g., Aberson 2008; Aberson et al. 2017). Motivated in part by the success of assimilating more reconnaissance data into HWRF, testing for the 2021 implementation of the GFS version 16 (GFSv16) included an assessment of adding more near-center dropwindsonde data as well as the first-time addition of flight-level HDOBs into the GFS. Preliminary analysis of the results led NCEP to include the additional data into GFSv16 (Farrar 2021a).

Here, a more comprehensive assessment of the impact of the additional data is presented. Section 2 describes the experiments, including the periods examined, the baseline configuration of GFS, and details regarding the additionally assimilated data. Sections 3–4 present results from different subsets of cases, and a summary and conclusions are given in section 5.

2. Methods

a. Experiment setup

This study uses a preoperational version of the NCEP GFSv16 that is nearly identical to the operational GFSv16. GFSv16 is based on GFS version 15 (GFSv15), with 13-km grid spacing and an upgraded physical parameterization package including GFDL microphysics (Zhou et al. 2019), an updated parameterization of ozone photochemistry with additional production and loss terms (McCormack et al. 2006), and a newly introduced parameterization of middle atmospheric water vapor photochemistry (McCormack et al. 2008). The data assimilation configuration in GFSv16 leverages the same GSI-based hybrid four-dimensional ensemble-variational (4DEnVar) solver (Kleist and Ide 2015) that was utilized in GFSv15. Changes in GFSv16 include increasing the vertical resolution from 64 to 127 levels and moving the model top to 80-km height, improved physics, using the gain form of the local ensemble transform Kalman filter (Lei et al. 2018), and employing the four-dimensional incremental analysis update technique for DA (Kleist et al. 2021; Yang et al. 2021; Lei and Whitaker 2016). Note that GFSv16 does not use any kind of vortex relocation or bogussing procedures, but it does assimilate minimum sea level pressure (Kleist 2011) from the TC vitals database (TCVitals).[1]

The two experiments here quantify the impact of assimilating additional reconnaissance data on TC forecasts from GFSv16, with the only difference between the experiments being that one experiment (OLD) does not assimilate the additional reconnaissance data, and the other experiment (NEW) does. Aside from some minor bug fixes that do not alter interpretation of results or applicability to GFSv16 (Farrar 2021b), NEW is identical to the operational GFSv16. All other operationally assimilated data in GFSv16 were included in both experiments. Note that there is still reconnaissance data assimilated into OLD (as with GFSv15), but only from dropwindsondes, and not all dropwindsonde data were assimilated. As part of operational preprocessing, dropwindsonde wind data in OLD were discarded within a critical radius from the center of any TC (which is the greater of 111 km or 3 times the radius of maximum wind speed) due to concerns regarding dropwindsonde drift (i.e., the lateral advection of dropwindsondes by the wind).[2] NEW relaxed these criteria for rejecting dropwindsonde wind data. In particular, all dropwindsonde data for TCs weaker than hurricane intensity in NEW were passed to the DA step, where it could be rejected depending on the difference between the data and first guess. In addition, for hurricanes the critical radius for the rejection of dropwindsonde wind data was reduced to 55 km with no consideration of the radius of maximum wind speed.[3] NEW

[1] TCVitals contains operational estimates of a TC's position, intensity, size, and motion (Trahan and Sparling 2012).

[2] Global DA at NCEP uses information only in the body of WMO TEMP DROP messages, which only contain the dropwindsonde release point to the nearest tenth of a degree. Thus, the entire dropwindsonde is assimilated in a column with a somewhat inaccurate initial location, and not considering the lateral drift. Since inner-core dropwindsondes in hurricanes sometimes travel to the opposite side of the storm (e.g., Aberson 2008), assimilating such wind data at the wrong location produces extreme, and erroneous analysis increments.

[3] These changes were made with consideration of how much drift was likely to be experienced by dropwindsondes in various wind conditions.

TABLE 2. Storms from 2018 verified in NEW against OLD. The dates of verification are given in the third column, and the periods of reconnaissance (if any) are shown in the final column.

Basin	Storm (ID)	Verification date range	Recon date range
NATL	Florence (06)	0600 UTC 2 Sep 2018–0600 UTC 17 Sep 2018	1200 UTC 8 Sep 2018–1800 UTC 14 Sep 2018
NATL	Gordon (07)	0600 UTC 3 Sep 2018–1200 UTC 6 Sep 2018	1200 UTC 3 Sep 2018–0600 UTC 4 Sep 2018
NATL	Helene (08)	1200 UTC 7 Sep 2018–0600 UTC 16 Sep 2018	—
NATL	Isaac (09)	1200 UTC 7 Sep 2018–0000 UTC 15 Sep 2018	1200 UTC 12 Sep 2018–0600 UTC 15 Sep 2018
NATL	Joyce (10)	0000 UTC 14 Sep 2018–1800 UTC 18 Sep 2018	—
EPAC	Miriam (15)	0600–1200 UTC 2 Sep 2018	—
EPAC	Norman (16)	0600 UTC 2 Sep 2018–1800 UTC 8 Sep 2018	1800 UTC 4 Sep 2018–0000 UTC 6 Sep 2018
EPAC	Olivia (17)	0000 UTC 1 Sep 2018–0000 UTC 14 Sep 2018	1800 UTC 8 Sep 2018–0000 UTC 12 Sep 2018
EPAC	Paul (18)	0600 UTC 8 Sep 2018–1800 UTC 11 Sep 2018	—

also added flight-level HDOBs, which are assimilated for the first time in GFS. While SFMR data are also included in HDOB transmission, that data have not been added in these experiments and is not currently assimilated in GFSv16. Hereafter in this study, HDOBs refer to only the flight-level HDOBs.

Settings for thinning and assumed observation errors were obtained from GFSv15 for dropwindsonde observations and from HWRF for HDOB data. The assumed observation error for both data types is shown in Table 1 for reference. For dropwindsondes released from the high-altitude G-IV, the number of vertical levels (i.e., mandatory and significant levels in WMO TEMP DROP data) in each dropwindsonde can exceed 50. For lower altitude reconnaissance, the number of levels in each dropwindsonde is typically less than 15. Thinning is not applied to either data type, and for HDOB this means the nominal data resolution is about 3 km along the flight track. We acknowledge that assimilating the HDOB data unthinned is suboptimal in GFSv16, and future results can likely improve with optimized assimilation parameters for both data types. Further, the specified observation error needs tuning for both HDOBs and dropwindsondes. Finally, the same temperature bias correction scheme was used for the HDOB data as other aircraft data (Zhu et al. 2015).

Given that the vast majority of reconnaissance missions sample North Atlantic basin (NATL; including the North Atlantic Ocean, the Gulf of Mexico, and the Caribbean Sea) TCs, the tests presented here focused on periods of NATL reconnaissance. The periods of interest for which NEW ran include the following: 1) 0000 UTC 1 September 2018–1800 UTC 18 September 2018, 2) 0600 UTC 22 August 2019–0600 UTC 2 October 2019, 3) 1800 UTC 1 June 2020–1800 UTC 9 June 2020, and 4) 0600 UTC 20 July 2020–1800 UTC 4 August 2020. The storms encompassed by these periods are listed in Tables 2–4 for 2018–20, respectively. These periods include the most active periods of the 2018–19 NATL seasons and the storms with NATL reconnaissance in 2020 that had occurred up until the point when the NEW experiments were conducted. Note that eastern North Pacific basin (EPAC) TCs occurred during some of these periods, and this study examines both the direct and remote impacts of the additional reconnaissance data on those TCs as well. Though there were also storms in the western North Pacific (WPAC), substantial sample size constraints limit interpretation of WPAC results. The entire WPAC sample contains less than 150 forecasts at 0 h, and it decreases to less than 25 by 72 h. Results were quite noisy and are not further reported below.

b. *Verification*

Verification is performed according to standard National Hurricane Center (NHC) procedures against the best tracks

TABLE 3. As in Table 2, but for 2019.

Basin	Storm (ID)	Verification date range	Recon date range
NATL	Dorian (05)	1200 UTC 24 Aug 2019–1200 UTC 7 Sep 2019	1800 UTC 25 Aug 2019–1800 UTC 6 Sep 2019
NATL	Erin (06)	1800 UTC 26 Aug 2019–0600 UTC 29 Aug 2019	—
NATL	Fernand (07)	1200 UTC 3 Sep 2019–0000 UTC 5 Sep 2019	1800 UTC 3 Sep 2019–0600 UTC 4 Sep 2019
NATL	Gabrielle (08)	1800 UTC 3 Sep 2019–0600 UTC 10 Sep 2019	—
NATL	Humberto (09)	1800 UTC 13 Sep 2019–1800 UTC 19 Sep 2019	1800 UTC 12 Sep 2019–0600 UTC 19 Sep 2019
NATL	Jerry (10)	1200 UTC 17 Sep 2019–0600 UTC 25 Sep 2019	1200 UTC 18 Sep 2019–0600 UTC 24 Sep 2019
NATL	Imelda (11)	1200 UTC 17 Sep 2019–1200 UTC 19 Sep 2019	—
NATL	Karen (12)	0600 UTC 22 Sep 2019–1200 UTC 27 Sep 2019	1800 UTC 21 Sep 2019–1800 UTC 26 Sep 2019
NATL	Lorenzo (13)	0000 UTC 23 Sep 2019–0600 UTC 2 Oct 2019	1200 UTC 27 Sep 2019–0000 UTC 29 Sep 2019
EPAC	Ivo (10)	0600 UTC 22 Aug 2019–1200 UTC 25 Aug 2019	1800 UTC 24 Aug 2019–0000 UTC 25 Aug 2019
EPAC	Juliette (11)	0600 UTC 1 Sep 2019–1200 UTC 7 Sep 2019	—
EPAC	Akoni (12)	1200 UTC 4 Sep 2019–0600 UTC 6 Sep 2019	—
EPAC	Kiko (13)	1800 UTC 12 Sep 2019–1800 UTC 24 Sep 2019	—
EPAC	Mario (14)	1200 UTC 17 Sep 2019–0000 UTC 23 Sep 2019	—
EPAC	Lorena (15)	1200 UTC 17 Sep 2019–0600 UTC 22 Sep 2019	1800 UTC 20 Sep 2019–0000 UTC 22 Sep 2019
EPAC	Narda (16)	0000 UTC 29 Sep 2019–0600 UTC 1 Oct 2019	—

TABLE 4. As in Table 2, but for 2020.

Basin	Storm (ID)	Verification date range	Recon date range
NATL	Cristobal (03)	1800 UTC 1 Jun 2020–1800 UTC 9 Jun 2020	1200 UTC 2 Jun 2020–1200 UTC 7 Jun 2020
NATL	Gonzalo (07)	1800 UTC 21 Jul 2020–1800 UTC 25 Jul 2020	1200 UTC 24 Jul 2020–0600 UTC 25 Jul 2020
NATL	Hanna (08)	0000 UTC 23 Jul 2020–0000 UTC 27 Jul 2020	1800 UTC 22 Jul 2020–1800 UTC 25 Jul 2020
NATL	Isaias (09)	0000 UTC 30 Jul 2020–1800 UTC 4 Aug 2020	1200 UTC 28 Jul 2020–0600 UTC 4 Aug 2020
NATL	Ten (10)	1800 UTC 31 Jul 2020–1800 UTC 1 Aug 2020	—
EPAC	Seven (07)	0600 UTC 20 Jul 2020–0600 UTC 21 Jul 2020	—
EPAC	Douglas (08)	0000 UTC 21 Jul 2020–0600 UTC 29 Jul 2020	—

available from NHC. Only tropical and subtropical systems of at least depression intensity are verified, and no verification is performed if the NCEP TC tracker (Marchok 2002) is unable to find a vortex associated with a given system. No verification is performed for situations with spurious TCs, either from false alarms, or when the model continues to forecast a TC beyond its dissipation in the best track. The statistics shown below are homogeneous, meaning that both OLD and NEW had to have a verifiable forecast for a given cycle in order for it to be included in the sample.

The one exception to using NHC verification rules is that the EPAC verification here includes storms initialized between 140°W and 180° longitude, which helps consolidate the limited sample of reconnaissance cycles in the EPAC. Operationally, NHC only verifies EPAC cycles if a storm is initialized east of 140°W, which is the limit of their area of responsibility (AOR). Using the expanded domain thus includes storms that operationally fell within the AOR of the Central Pacific Hurricane Center (CPHC).

For completeness, we verify both "early" and "late" model forecasts in this study. The late model forecast is the raw output from the vortex tracker, but for a given cycle it is typically not available until after operational centers issue their forecasts. Thus, operational centers typically base their forecasts off postprocessed output from the most recent available cycle. The postprocessing procedure, otherwise known as the interpolator, applies a bias correction determined by the bias in the previous 6-h forecast for position and intensity. The interpolator applies the same position correction through the duration of the forecast, whereas the intensity correction linearly decreases to 0 at a given lead time (120 h in GFSv16). In addition to bias correction, 10 passes of a 1:2:1 smoother are run over the tracker 6-h intensity output for the early forecast. The bias correction procedures are the reason why track and intensity errors are near 0 at 0 h in the early-model forecasts in sections 3–4.

Verification metrics used here include the mean absolute error (MAE; the standard NHC verification metric), the median absolute error, mean error (i.e., bias), and the percent improvement (i.e., skill) of both the MAE and the median. As in Rappaport et al. (2009), track error is defined as the great-circle distance between the forecast and best track positions, and intensity error is defined as the difference in the observed and tracker-calculated maximum sustained 10-m wind speed (i.e., Vmax). Though median errors are not typically verified by NHC, we believe that they relay important information not captured by the MAE since MAE can be strongly influenced by the tail of the distribution. At times the two metrics differ considerably.

Finally, we test for significant differences in MAE between the two experiments using a paired-sample t test, which is the appropriate test since track and intensity errors in OLD and NEW are well correlated at most verification times through day 7 (not shown). While testing for significance, we account for serial error correlation in the following process:

- Calculate the time-lag forecast error correlation in NEW for each variable at each lead time out to as many as five cycles, or 24 h;

TABLE 5. Reduction factors from the adaptive serial correlation used for calculating statistical significance in each basin as a function of lead time. Factors are given individually for track (TK) and Vmax (VM) for the NATL (AL) and EPAC (EP) according to lead time (in days).

Day	0				1				2				3				4				5				6				7	
LATE																														
AL—TK	2	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
AL—VM	3	3	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	4	5	3	3	2
EP—TK	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	2	2	1	2	1
EP—VM	2	2	2	2	2	2	2	3	3	3	3	4	4	4	4	4	3	3	3	3	2	2	2	3	3	3	3	3	3	3
EARLY																														
AL—TK	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
AL—VM	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	5	5	5	5	5	5	5	4	3	3	2	
EP—TK	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	4	3	3	2	2	1	2	1	1	
EP—VM	1	1	2	2	2	2	2	3	3	3	3	4	4	4	4	4	4	3	3	3	2	3	3	3	3	3	3	3	2	

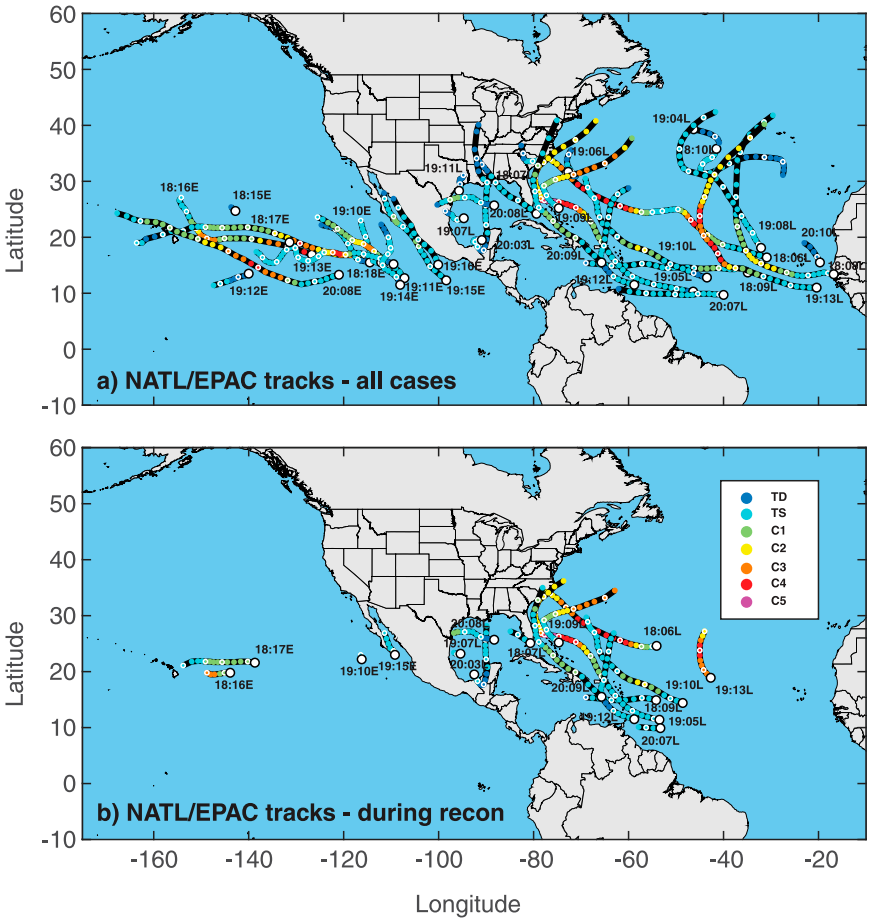

FIG. 1. The portions of tracks of TCs included in this study for which forecasts have been verified. Tracks are shown for (a) all cases and (b) only during periods of reconnaissance, defined as the first to last cycle when reconnaissance occurred for a given storm. The initial position in each track is marked with a white circle with a black outline, and each storm is identified according to its year and storm number (e.g., 19:05L corresponds to NATL storm 05 from the 2019 season). Further, the intensity at each location is color coded according to the legend, and white circle outlines along each track indicate each 0000 UTC time.

- Assign a reduction factor to be the number of cycles (up to 5) to be used in all subsequent comparisons that it takes for the error correlation to decrease to less than ~0.7 (i.e., the explained variance decreases to less than 50%);
- Divide the total sample size at each lead time by the corresponding reduction factor, thus creating an effective sample size used for significance testing.

Table 5 shows the reduction factors for the early and late track and intensity forecasts for each basin. Note that the factors are quite different for track and intensity and from basin to basin. This "adaptive serial correlation" approach differs from recent studies (e.g., Alaka et al. 2017) that have used constant reduction factors for all lead times and metrics, but we believe it more accurately accounts for nuances in the error statistics. Finally, lead times when MAE in OLD and NEW are statistically different with 95% confidence are marked in each figure.

c. *Reconnaissance sampling*

Figure 1 gives a sense of where TCs in this study were sampled and at what point in their life cycle they were sampled. Note that TC locations included were only those that would verify under the standard NHC verification rules as detailed in section 2b. Figure 1a depicts the tracks of all NATL and EPAC TCs within the periods of interest (Tables 2–4, third column). Meanwhile, Fig. 1b shows the tracks of the same storms but only during periods of reconnaissance (Tables 2–4, fourth column).

The majority of reconnaissance focused on storms in either the main development region or the subtropical Atlantic that posed threats to either Caribbean islands or the continental United States. There were also flights into some shorter-lived storms in the Gulf of Mexico and some brief periods of coverage in the EPAC. Reconnaissance often begins during the early and weaker stages of storms and continues as they

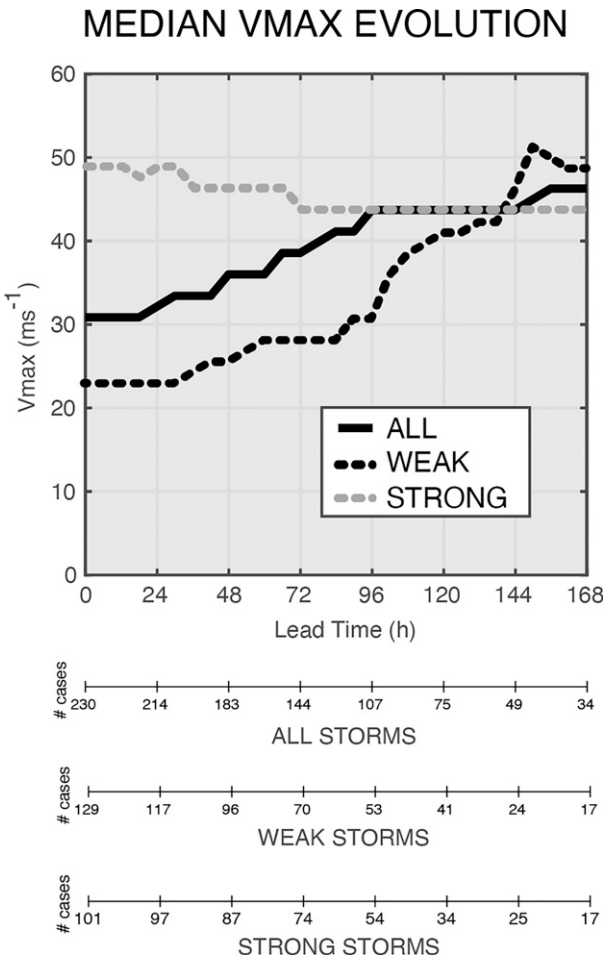

FIG. 2. The median Vmax evolution, calculated as described in section 2c, for all (solid), initially weak (solid dashed), and initially strong (gray dashed) storms.

strengthen, but that is not always the case. For example, missions began with Isaac (18:09L in Fig. 1) as it was decaying and headed toward the Caribbean islands.

To further illustrate how reconnaissance missions related to the evolution of observed storm intensity in a bulk sense, we calculate a median observed intensity for the reconnaissance periods. To make this calculation, we identify the first cycle with reconnaissance data for a given TC. Corresponding with this cycle, the observed Vmax value in the best track is assigned a "lead time" of 0 h (i.e., the cycle time), and subsequent observed Vmax values are assigned "lead times" through 168 h later. This process is repeated for subsequent cycles through the last cycle with reconnaissance, thereby creating a sample of the observed Vmax evolution. Note that this sample includes the portions of storms identified in Fig. 1b and in the fourth column of Tables 2–4. Figure 2 depicts the median Vmax evolution of all TCs shown in Fig. 1b (solid line) as well as the median Vmax evolution for those TCs in Fig. 1b classified at 0 h of each cycle as weak storms (tropical depressions and tropical storms; dashed black line) and strong storms (hurricanes; dashed gray line).

The Vmax evolution of storms in this study during and after periods of reconnaissance varies strongly depending on the initial storm intensity. The typical reconnaissance mission sampled a strong tropical storm, and if the system remained tropical, it tended to be stronger over the subsequent 7 days (Fig. 2, solid line). The stratification between strong and weak systems reveals that missions flew into weak storms more frequently (e.g., 129 versus 101 cases at 0 h), though the sample size of the two groups was similar after the first day due to a larger attrition rate of weak systems. Focusing just on the weak storms that were sampled, the systems that remained tropical tended to strengthen a great deal after a week. On day 7, the median intensity for initially weak storms was around the major hurricane threshold. Meanwhile, strong storms (i.e., hurricanes) sampled by reconnaissance missions tended to weaken slightly over the subsequent week if they remained tropical.

d. *Example distribution of new data*

The additional HDOB data accounts for the vast majority of added data. To help illustrate this point, Fig. 3 compares the distribution of reconnaissance wind data assimilated in OLD and NEW during Hurricane Dorian (2019). In total, NEW assimilated over 37 000 HDOB wind observations during Dorian, while OLD did not assimilate any HDOB data (Figs. 3a,c). Aside from extensive HDOB data in and around Dorian, a large amount of data lies to the west along the flight transit paths, which originate and terminate in either the country of Aruba for the flights farthest east, or in Lakeland, Florida, or Biloxi, Mississippi, as the storm neared the United States (there was also a single mission into Tropical Storm Fernand in the Gulf of Mexico during this period). In a single well-sampled cycle near Dorian's peak intensity (Fig. 3e), NEW assimilated nearly 1300 HDOB wind observations.

As designed, NEW also assimilated a number of additional dropwindsonde wind observations near Dorian's center (Fig. 3, right column). The difference in the number of dropwindsonde wind observations assimilated between OLD and NEW was roughly 2500, which is an approximate 25% increase in NEW (Figs. 3b,d). The single aforementioned well-sampled cycle assimilated about 70 more dropwindsonde wind observations in NEW than OLD (Fig. 3f), concentrated near the 75-km radius. In total, NEW assimilated almost 13 000 dropwindsonde wind observations over Dorian's lifetime, about a third the amount of the additional HDOB wind data assimilated in the same period.

3. NATL results

This section presents the impact of assimilating additional reconnaissance data in NEW on various aspects of NATL TC forecasts. Considering the primary interest of assimilating additional reconnaissance data is to improve the forecasts of storms with reconnaissance, we first present results focusing exclusively on storms with reconnaissance from the first to the last cycle with reconnaissance data (Figs. 4–10; Tables 2–4, fourth column). Note that there are occasionally cycles without reconnaissance in these periods, though errors are typically well correlated for at least a cycle (e.g., Table 5). [Figures 11–13 then examine impacts on the full NATL

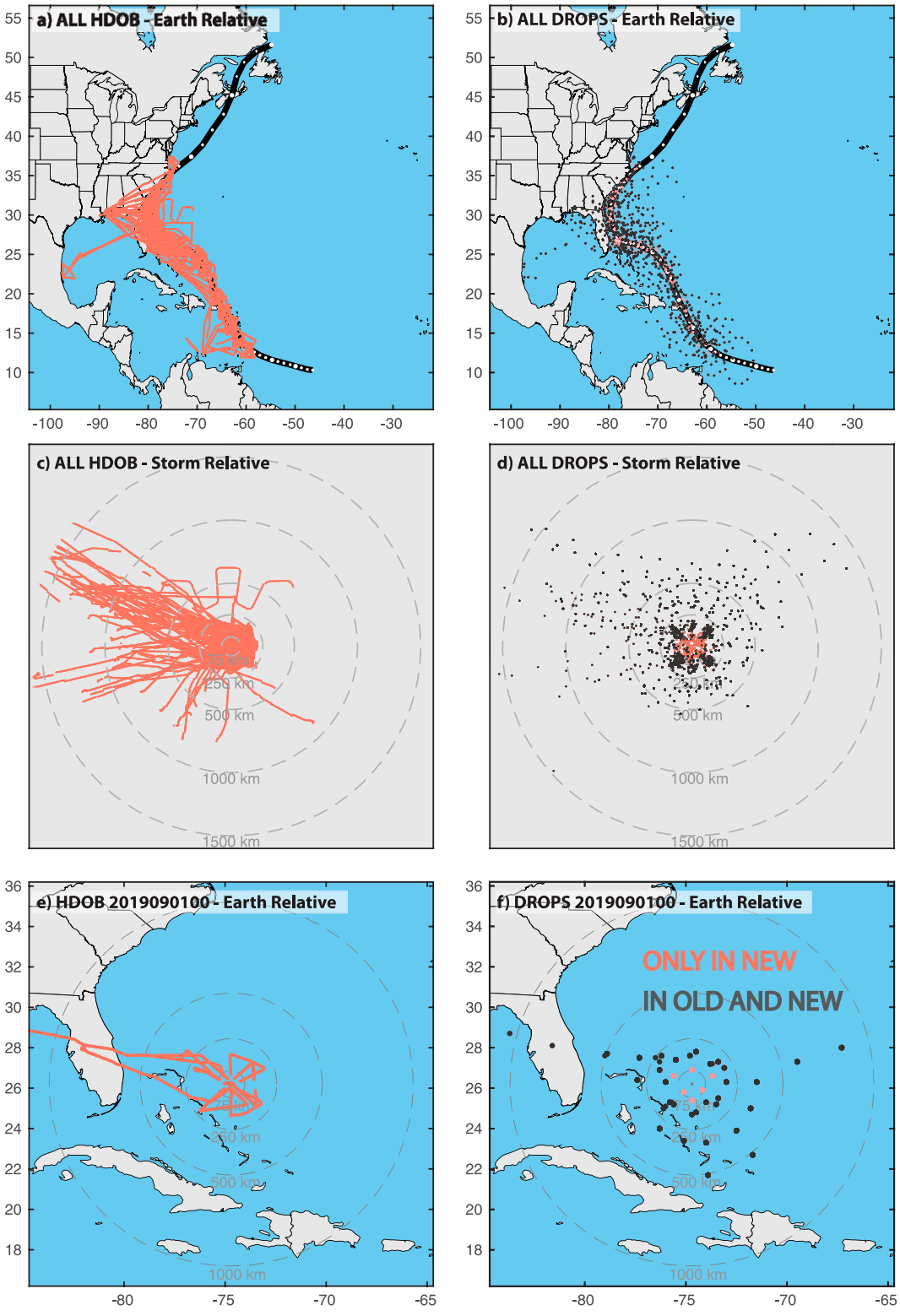

FIG. 3. Assimilated wind data from (left) HDOBs and (right) dropwindsondes (DROPS) during Hurricane Dorian (a),(b) for the duration of the storm (0600 UTC 24 Aug 2019–0000 UTC 9 Sep 2019) in a regional, Earth-relative perspective; (c),(d) for the duration of the storm in a storm-centered perspective; and (e),(f) during the 0000 UTC 1 Sep 2019 cycle. Note that the gray markers indicate observations assimilated in both NEW and OLD, while the coral markers indicate observations only assimilated in NEW.

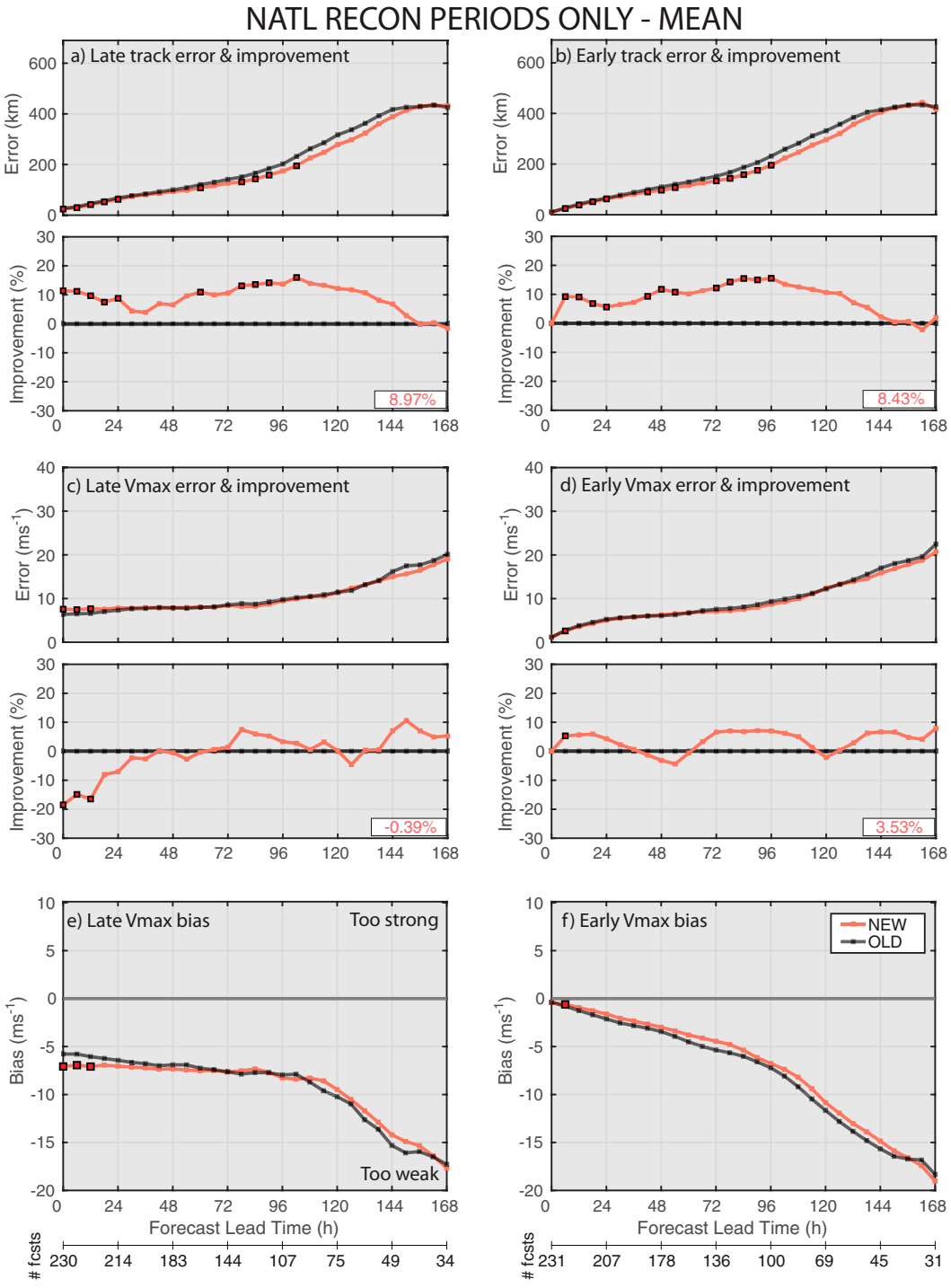

FIG. 4. Track and intensity verification evaluated only for the periods of reconnaissance in NATL storms that had reconnaissance (i.e., the final column in Tables 2–4) in terms of (a)–(d) MAE and percent improvement of the MAE and (e)–(f) intensity bias. (left) Late verification and (right) early verification, and the number of cases at each lead time is shown at the bottom. The average percent improvement across all lead times is shown in the bottom right of (a)–(d). Markers indicate lead times where mean errors are statistically different at the 95% confidence level based on a paired *t* test with adaptive serial correlation.

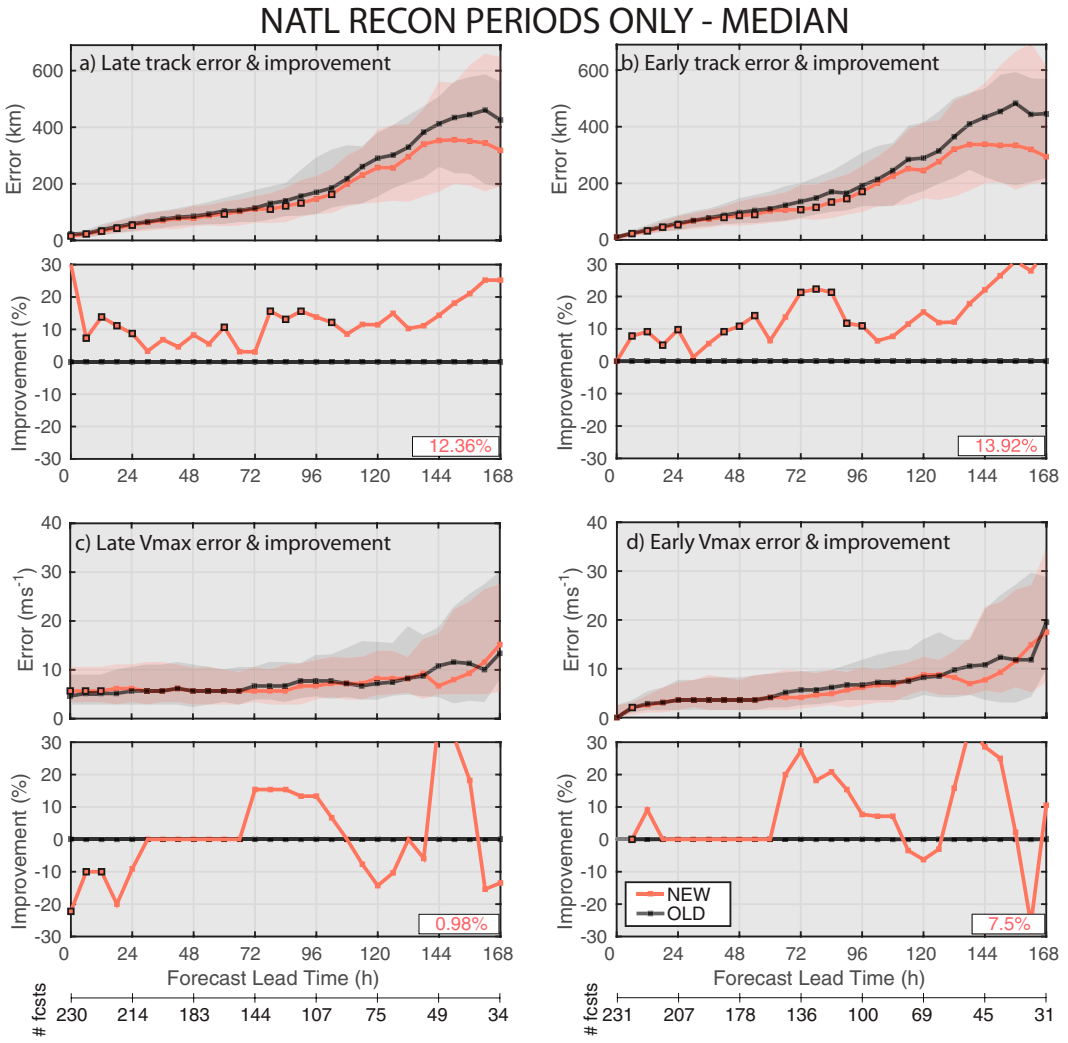

FIG. 5. As in Figs. 4a–d, but for median absolute errors and percent improvement of the median error. Lead times with statistically significant differences in Fig. 4 are also indicated here for reference. The IQR of each experiment is shown with transparent shading.

sample, including cases without reconnaissance (Tables 2–4, third column).]

a. NATL reconnaissance periods

The most important result from this study is that including additional reconnaissance data in NEW significantly reduces both mean and median track errors in the NATL storms being sampled. In Fig. 4, which shows verification in terms of MAE and percent improvement for NATL reconnaissance storms during sampling periods, track improves by about 9% (8.4%) on average, and by as much as 16% (15.5%) at individual lead times in late (early) model forecasts when the additional data are assimilated (Figs. 4a,b). The improvement is statistically significant at many lead times through day 4, and it remains substantial though not statistically significant after that. On average, the additional data reduces track error on days 4–5 by 35–40 km and results in superior forecasts at almost every lead time through day 7. Overall, the reduction in track error in NEW is associated with improved along track bias. Storms in both OLD and NEW tend to move faster than those in the best track, but the additional reconnaissance data substantially improves this trend in NEW at most verification times (not shown). The cross-track bias in this sample is fairly small, and the forecasts in OLD and NEW do not meaningfully differ in that respect (not shown).

The commensurate median track errors and interquartile range (IQR) in Figs. 5a and 5b give more insight into the error distributions in NEW and OLD. Consistent with the MAE, the median error and IQR bounds in NEW are smaller than in OLD through most of the first 5 days. On days 6–7 the median improvement in NEW remains quite large (>20%), and the lower bound of the IQR is equal to or lower than that in OLD. However, the upper bound of the IQR increases

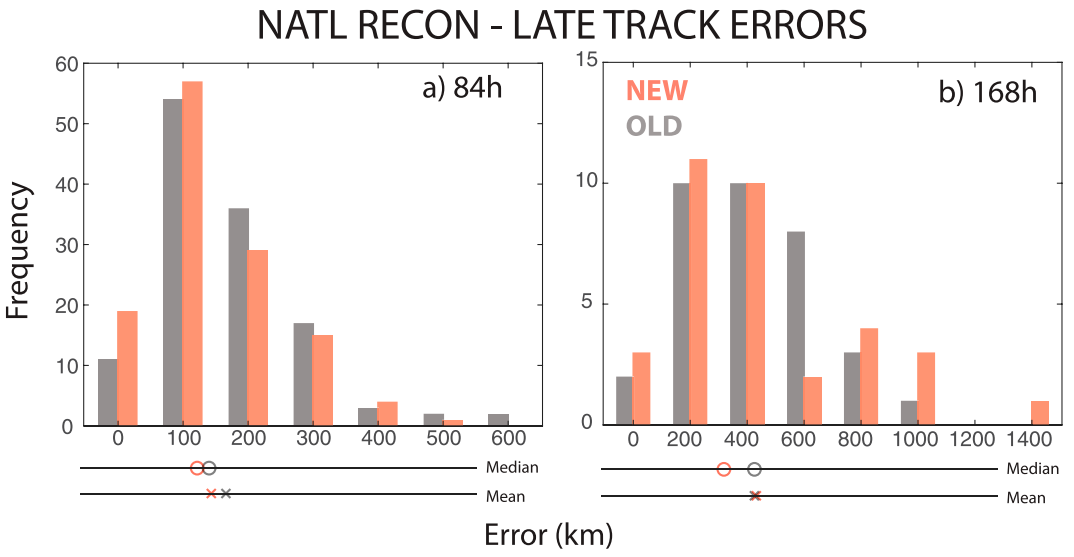

FIG. 6. Histograms of NATL absolute track errors in the recon subsamples of OLD and NEW at (a) 84 and (b) 168 h. The mean and median of the distributions are indicated below each histogram.

quite a bit in NEW during this time, suggesting an increasingly asymmetric distribution of errors about the median. This suggests the significance-test assumption of an approximately normal error distribution is poor and highlights the importance of evaluating not just the MAE but also the median error.

Histograms of late model track error at 84 and 168 h in Fig. 6 give further insight into how OLD and NEW forecasts behave. At 84 h (Fig. 6a), errors in both experiments have roughly a chi-squared distribution. NEW has more frequent errors in the bins < 200 km than OLD, whereas OLD has more frequent errors in nearly every bin ≥ 200 km than NEW. This uniform behavior yields similar improvements for NEW as assessed either through the median or mean. The distribution in NEW at 168 h is not as clean as at 84 h, possibly due to a limited sample size (~30 cases), which yields inconsistencies between the mean and median. NEW continues to have more frequent errors in the lower bins than OLD, but it also has a secondary frequency maximum around 800–1000 km. The secondary maximum in NEW along with an outlier of around 1400 km degrades its MAE relative to OLD but does not substantially impact the median (Fig. 6b).

The impact of additional reconnaissance data on NATL intensity forecasts is less straightforward than for track and depends on postprocessing procedures. In the late model, the additional data significantly degrades the intensity forecast for the first 24 h (Fig. 4c) due to negative intensity bias (i.e., storms in NEW are too weak, Fig. 4e). The interpolator corrects the negative intensity bias such that the early model intensity improves at most lead times when the reconnaissance data are assimilated (Fig. 4d). Overall, early intensity forecasts improve by 3.5% on average and up to 9%, though the improvement is not significant. The median errors reveal even greater improvement in NEW intensity at the extended lead times where the sample is small (Figs. 5 c,d). This again highlights the importance of evaluating not just the MAE but also the median error.

The additional data also improves some other aspects of TC forecasts, though those results are not shown for brevity. For example, forecasts of sea level pressure improve in NEW by up to 10%–20% after about 72 h in both the mean and median. Furthermore, forecasts of the storm-force wind radius (i.e., 25 m s^{-1}, otherwise known as the 50-kt wind radius in NHC verification parlance) in NEW also improve upon OLD by 5%–10% at many verification times through day 6. Meanwhile, consistent improvement in the mean and median is not seen for other significant wind radii.

b. *Dependence of NATL results on initial intensity classification*

Here we examine how the impacts of the additional data evolve as the storms themselves evolve. To better illustrate this, Figs. 7–9, respectively, show track errors, intensity errors, and intensity bias for weak (tropical depression and tropical storm) and strong (hurricane) storms. For brevity, these results are only shown for the late model. Though the differences are interesting, care should be taken not to overgeneralize because the sample sizes are quite small beyond the first few days.

The track results detailed in section 3a have a fairly strong dependence upon the initial storm classification. One obvious difference between weak and strong systems in Figs. 7–8 is that the track errors tend to be larger for weak systems, a trend also seen in operational NHC forecasts (e.g., Fig. 4 of Cangialosi 2022). For both classifications of TCs, the impact of the additional data on the track forecast is on average about the same as in the non-stratified reconnaissance sample through day 4 (cf. Figs. 7–8a,b to Fig. 4a). One difference is that track improvement for sampled hurricanes in NEW is

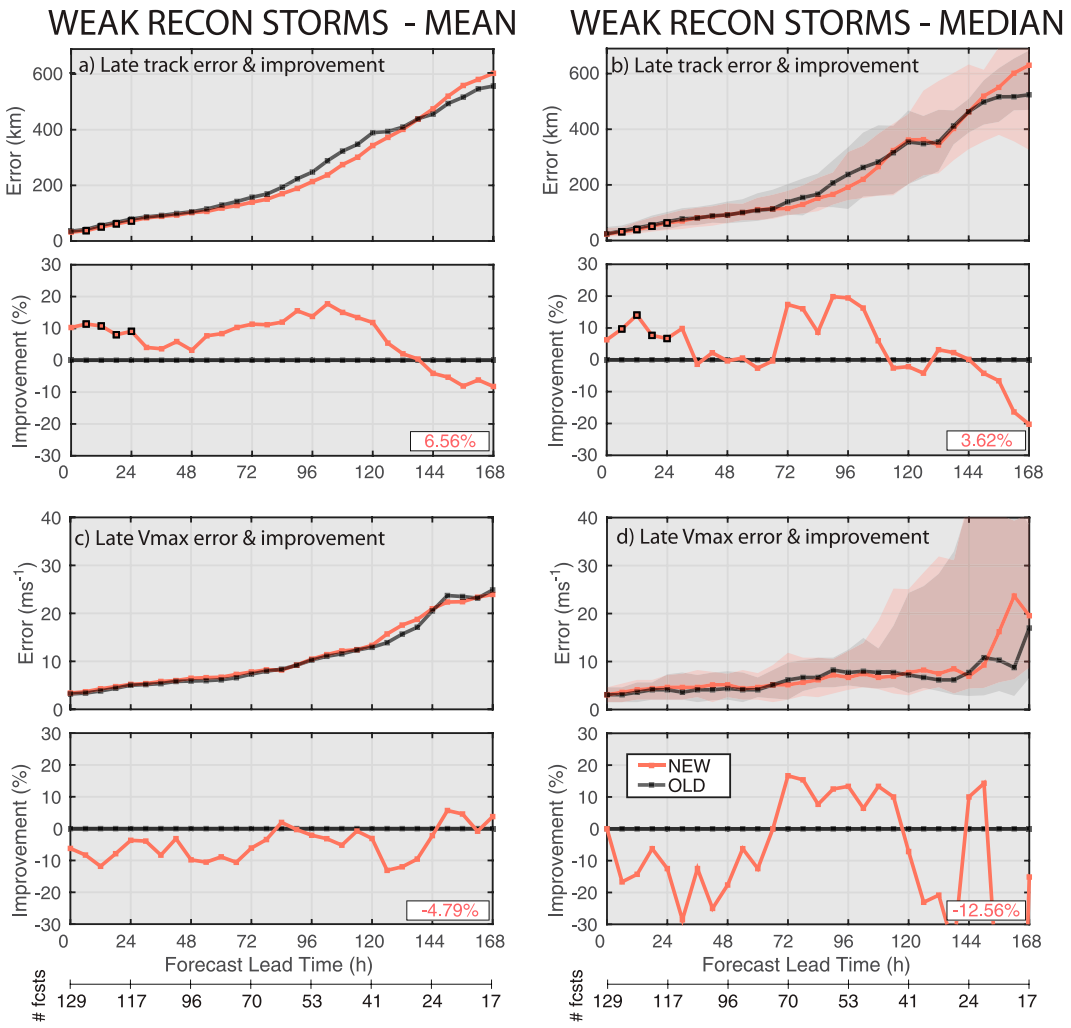

FIG. 7. Track and intensity verification of weak TCs (tropical depressions and tropical storms) from the sample in Figs. 4–5 in terms of (a),(c) MAE and percent improvement of the MAE and (b),(d) median errors and percent improvement of the median error. Lead times with statistically significant differences in MAE are indicated. Only late verification is shown, and the figure is otherwise identical to previous figures.

statistically significant on days 2–3, but differences for weak systems are not significant beyond day 1. Meanwhile, on days 5–7 the track error for weak systems in NEW is worse than OLD in terms of both the MAE and median absolute error (Figs. 7a,b). The results are much different for hurricanes, where MAE improvement in NEW is statistically significant on day 6 and exceeds 20% at several lead times on days 6–7 (Fig. 8a). Additionally, the median improvement often exceeds 20% during this period (Fig. 8b).

As with track errors, intensity forecast errors also depend strongly on the initial classification. Intensity errors for initially weak storms are slightly worse in NEW than in OLD at almost every verification time (Figs. 7c,d). Commensurate with these errors, the mean bias in NEW is consistently more negative than in OLD (Fig. 9a). In other words, if reconnaissance samples a weak system, the additional data tends to make the subsequent intensity forecast weaker and worse through the duration of the forecast. The results substantially differ when reconnaissance samples a hurricane. The additional data clearly makes the subsequent intensity forecast in NEW much weaker than in OLD for the first 36 h (Fig. 9b), which explains why the initial late-model intensity errors are much worse in NEW for strong storms (Figs. 8c,d). After 36 h the forecast intensity is on average stronger in NEW than OLD (Fig. 9b), and the improved bias in NEW is a substantial reason why the intensity MAE and median errors generally also improve (Figs. 8c,d).

The large degradation in the short-term late intensity forecast for hurricanes in NEW reflects a known issue when assimilating inner-core data in TCs. The substantial negative initial intensity bias imparted by the additional reconnaissance data is a much larger problem for hurricanes than for weak systems (Fig. 9), and it clearly drives the negative bias and larger absolute errors in Figs. 4–5. This result is similar to

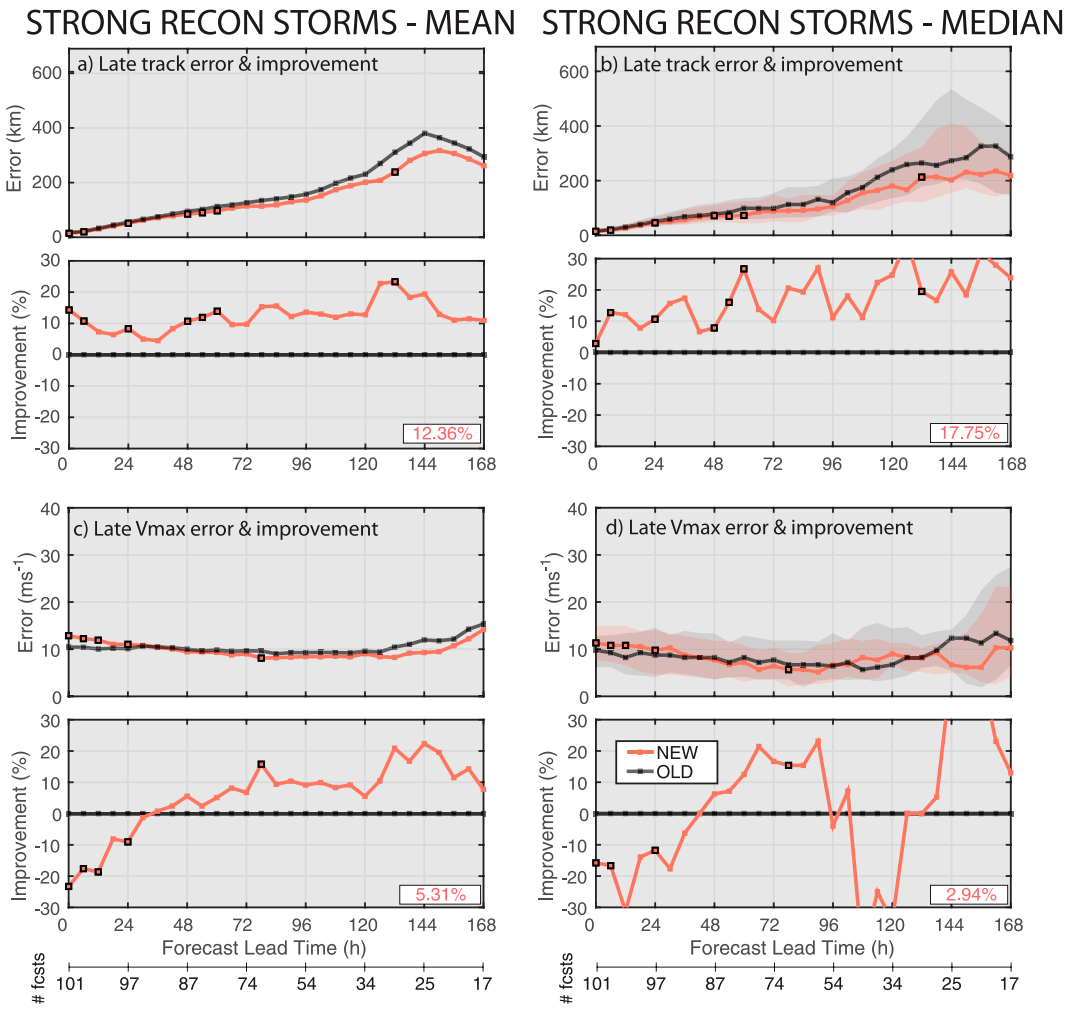

FIG. 8. As in Fig. 7, but for strong TCs (i.e., hurricanes).

those in Tong et al. (2018), who found that assimilating inner-core data in HWRF caused serious degradations to the short-term intensity forecast (their Fig. 3). The degradation in that paper was likewise due to large negative intensity biases that resulted from assimilating inner-core reconnaissance data into hurricanes (their Fig. 4). In their case, the negative intensity biases were due to model physics deficiencies and inappropriate error covariances in the hurricane core. In the present case, the current GFSv16 grid spacing (13 km) is clearly insufficient to resolve the inner-core structure of a hurricane, so the late-model intensity degradation is not particularly surprising.

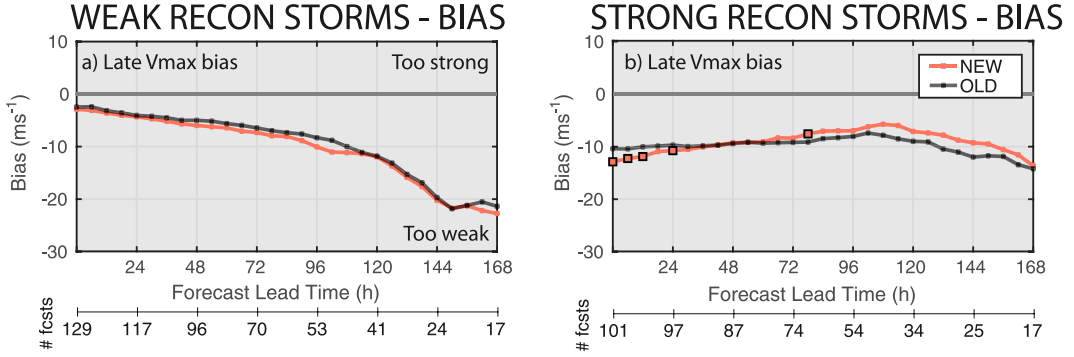

FIG. 9. As in Fig. 4e, but for (a) weak and (b) strong TCs.

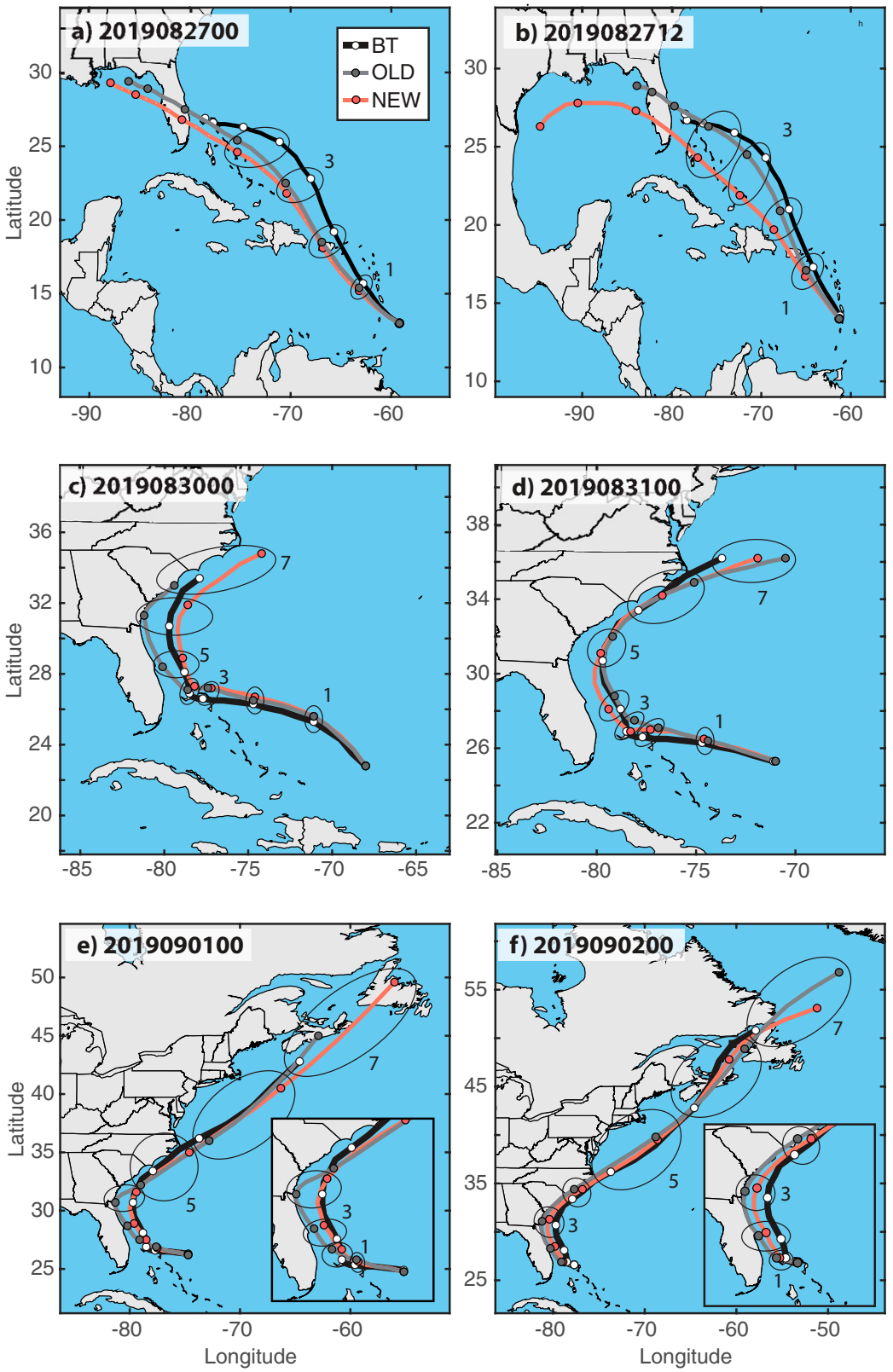

FIG. 10. (a)–(f) A comparison of early-model track forecasts from select cycles of Hurricane Dorian. Where practical, ellipses encircle the best track and forecast locations valid at the same time. To facilitate interpretation, odd verification days (e.g., 1, 3, 5, 7) are noted on the appropriate ellipses, and zoom-in panels are added in (e) and (f).

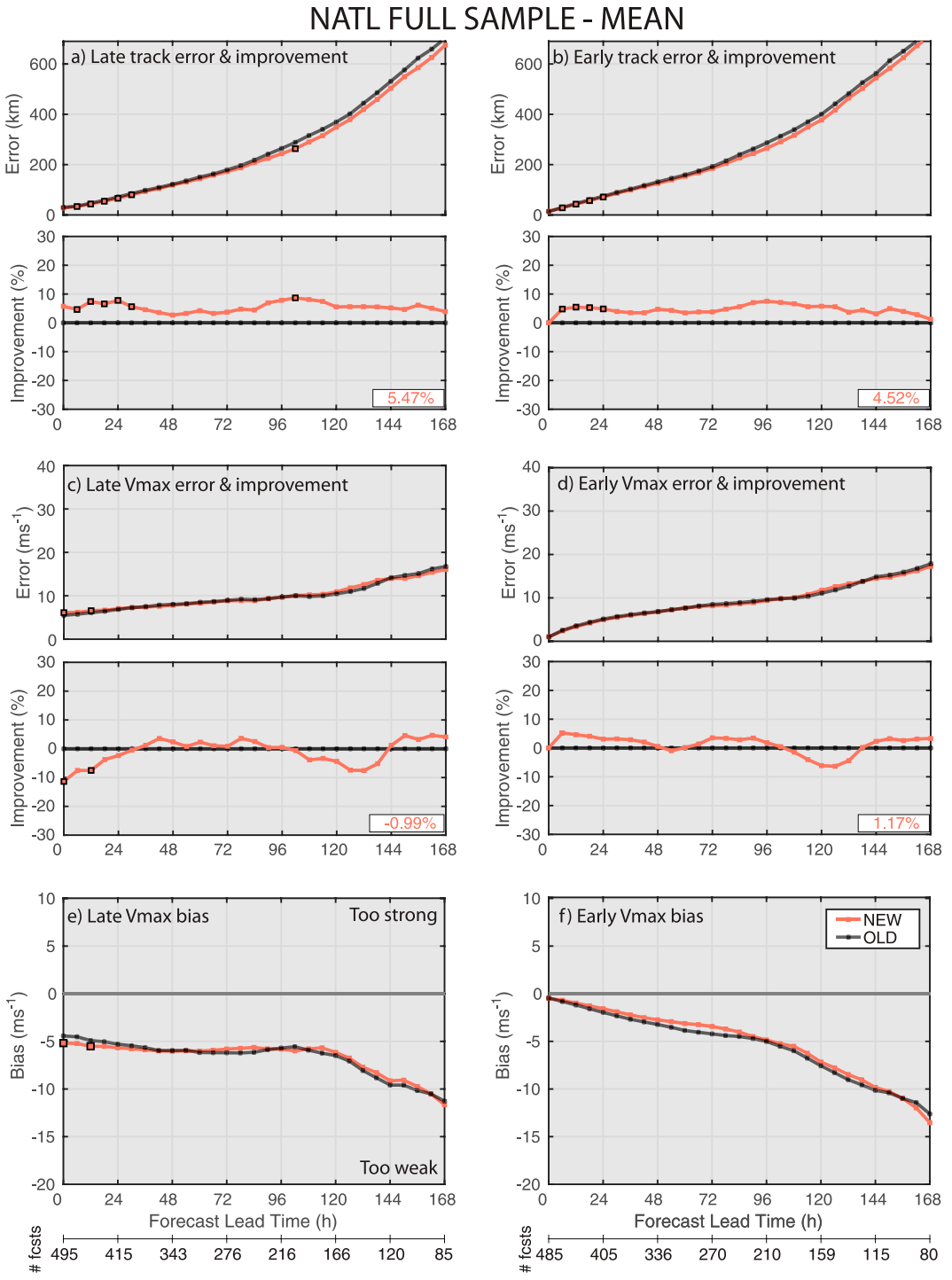

FIG. 11. As in Fig. 4, but for the entire NATL sample.

While the above results give an overall sense of the situationally dependent error characteristics, it is also useful to see the impacts of the data in individual cases. As such, Fig. 10 explores the differences between OLD and NEW over a series of forecasts of Hurricane Dorian as it approached the Southeast United States. The first two forecasts (Figs. 10a,b) encompass the period when Dorian was a disorganized tropical storm and inner-core reconnaissance was just beginning. The subsequent four forecasts (Figs. 10c–f) encompass the time when Dorian was near peak intensity, including the 0000 UTC 1 September 2019 cycle shown in Figs. 3e and 3f. For brevity, the focus here is exclusively on the track forecast.

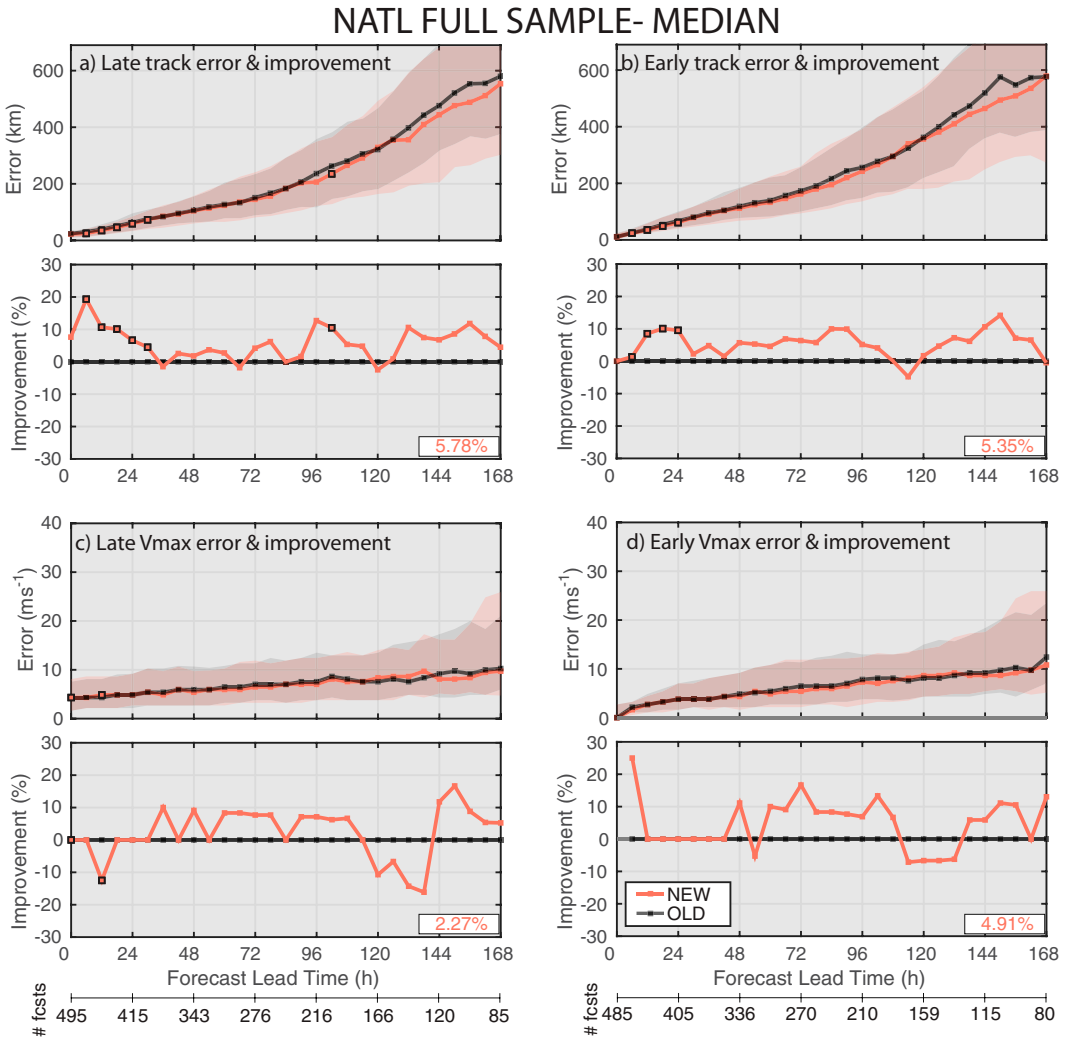

FIG. 12. As in Fig. 5, but for the entire NATL sample.

There is a very clear disparity in forecast errors in NEW from the early to mature stages of the storm. For the two weaker cycles in Fig. 10, the track forecasts in NEW are somewhat to significantly worse than those in OLD. In fact, the 1400-km error from Fig. 6b is from the forecast in Fig. 10b at 1200 UTC 27 August 2019. Subsequent track errors in NEW become competitive with or superior to those in OLD within another day as Dorian moves northwest and strengthens (not shown). For the mature stage, one can immediately see that the NEW forecasts tend to lie to the right of the OLD forecasts during the period of recurvature, which better agrees with the best track. In particular, three of the four OLD forecasts during this period show landfall in Florida, Georgia, or South Carolina, whereas the NEW forecasts correctly remain offshore of those states. Though the NEW forecasts are undoubtedly superior during Dorian's recurvature, the benefits of the additional reconnaissance data do not extend to the storm's northeastward acceleration.

c. *Entire NATL*

Results from the entire NATL sample, including periods and storms without reconnaissance, are qualitatively similar to the sample that focuses on reconnaissance cases. Track forecasts in NEW improve upon OLD by about 5.5% (4.5%) on average, and by as much as 8.7% (7.5%) at individual lead times in late (early) forecasts (Figs. 11a,b). Though including cycles without the additional data diminishes the impact, the improvement remains statistically significant during the first 36 h and on days 4–5. Similar to the smaller sample of cases with reconnaissance data, the late intensity forecasts from NEW significantly degrade during the first 12 h (Fig. 11c), again due to negative bias (Fig. 11e). As seen in the cases with reconnaissance, the interpolated intensity forecasts in NEW are somewhat better than those in OLD, particularly within the first 36 h (Fig. 11d). In terms of median errors, the track improvement in NEW is similar to the MAE improvement (Figs. 12a,b), though the intensity improvement is somewhat

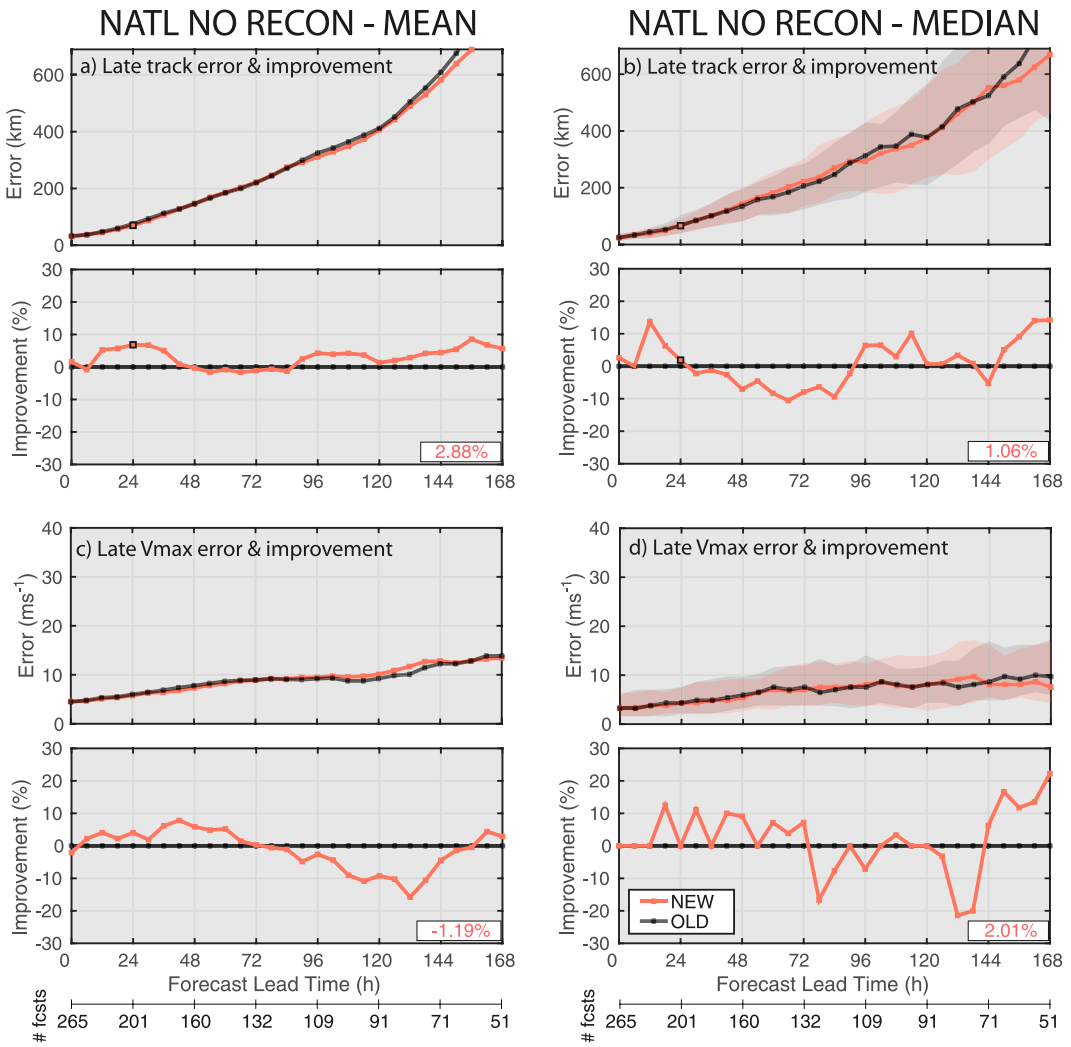

FIG. 13. Track and intensity verification during non-reconnaissance periods in terms of (a),(c) MAE and percent improvement of the MAE and (b),(d) median errors and percent improvement of the median error. Points of statistical significance from the MAE verification are indicated for reference. Only late verification is shown, and the figure is otherwise identical to previous figures.

larger after 48 h in the early model (Figs. 12c,d). A notable difference between OLD and NEW in terms of the track error IQR is that the lower bound of the IQR is quite a bit lower in NEW on days 6–7 (Figs. 12a,b). In summary, assimilating additional reconnaissance data improves even the entire NATL sample, especially in terms of the early model used in real time by NHC.

Considering that the number of cases outside reconnaissance periods constitutes roughly half the sample in Figs. 11–12, it is valuable to understand precisely how those cases contribute to the full-sample statistics. To do this, Fig. 13 shows the late model (early omitted for brevity) NATL track and intensity MAE and median errors and associated improvement for the non-reconnaissance portion of the sample. The NEW track MAE in Fig. 13a remains smaller than that in OLD, though the improvement diminishes in terms of the median (Fig. 13b). At the longer lead times, the MAE and median error both improve in NEW, and the NEW IQR tends to encompass lower errors, which suggests a small positive remote impact. Meanwhile, both experiments exhibit overall similar intensity errors (Figs. 13c,d). These results show that the diminished data impact in the full sample generally reflects a simple dilution and not negative remote impacts.

Curiously, a comparison of Fig. 13 with Figs. 4–5 reveals very different track error characteristics in the reconnaissance and no-reconnaissance samples. For example, in both OLD and NEW, median track errors remain less than ~300 km in the reconnaissance sample through 120 h. However, in the cases with only remote reconnaissance influence, the median track error at 120 h is 400 km. Such stark differences in error occur at most verification times for both the MAE and median

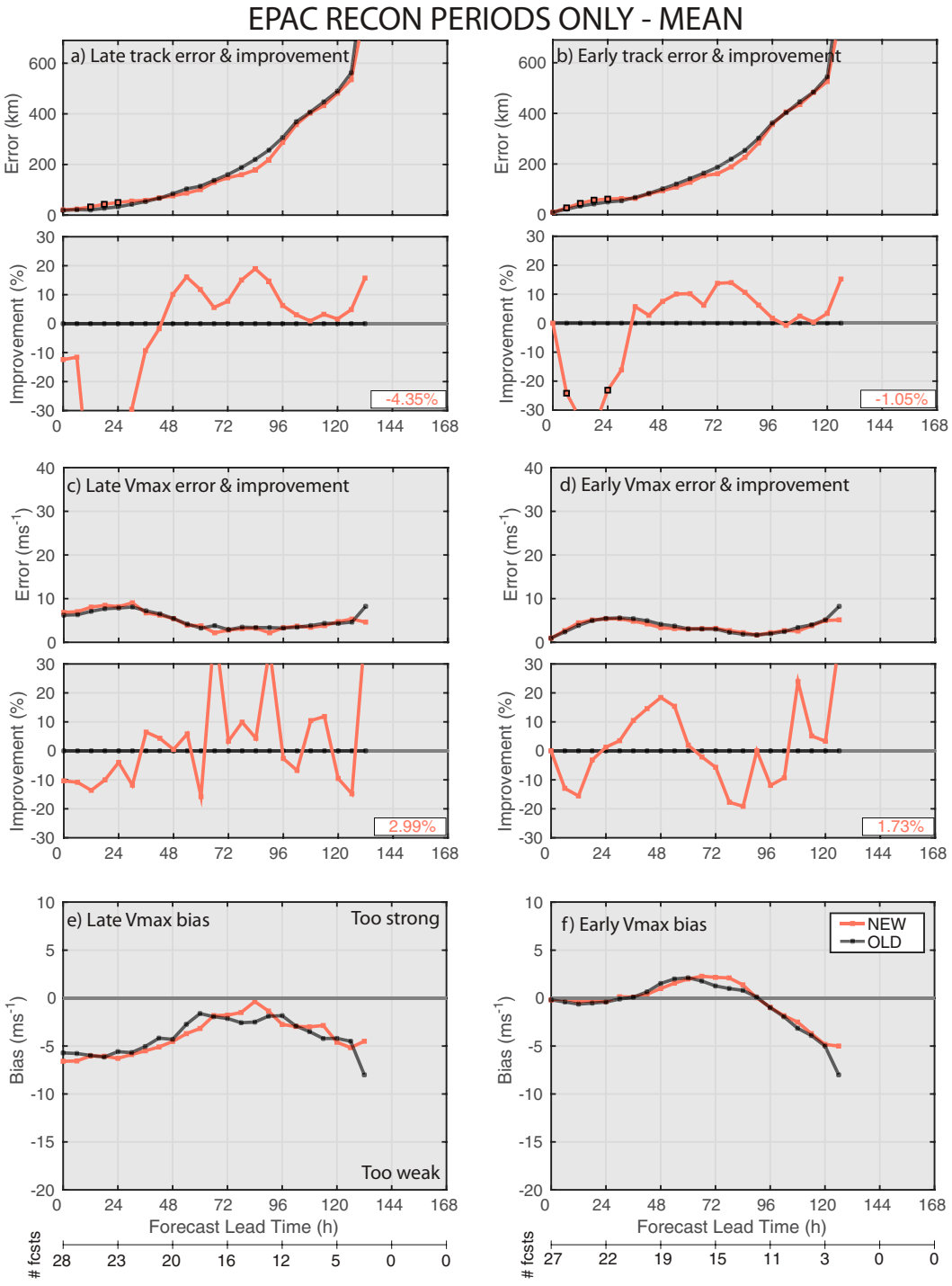

FIG. 14. As in Fig. 4, but for EPAC reconnaissance cases.

error. Two likely reasons for this difference are that the sample within reconnaissance periods contains other reconnaissance data (i.e., OLD and NEW treat environmental dropwindsondes identically) and that storms in reconnaissance periods tend to be closer to the data-rich U.S. observation network. In fact, internal testing at NOAA has shown that launching supplemental rawinsondes at 0600 and 1800 UTC in the U.S. network substantially reduces TC track errors (M. Brennan 2022, personal communication). Meanwhile, intensity errors in the reconnaissance period are not meaningfully different than in periods without reconnaissance. Though these comparisons are anecdotal since they cover different storms, they are

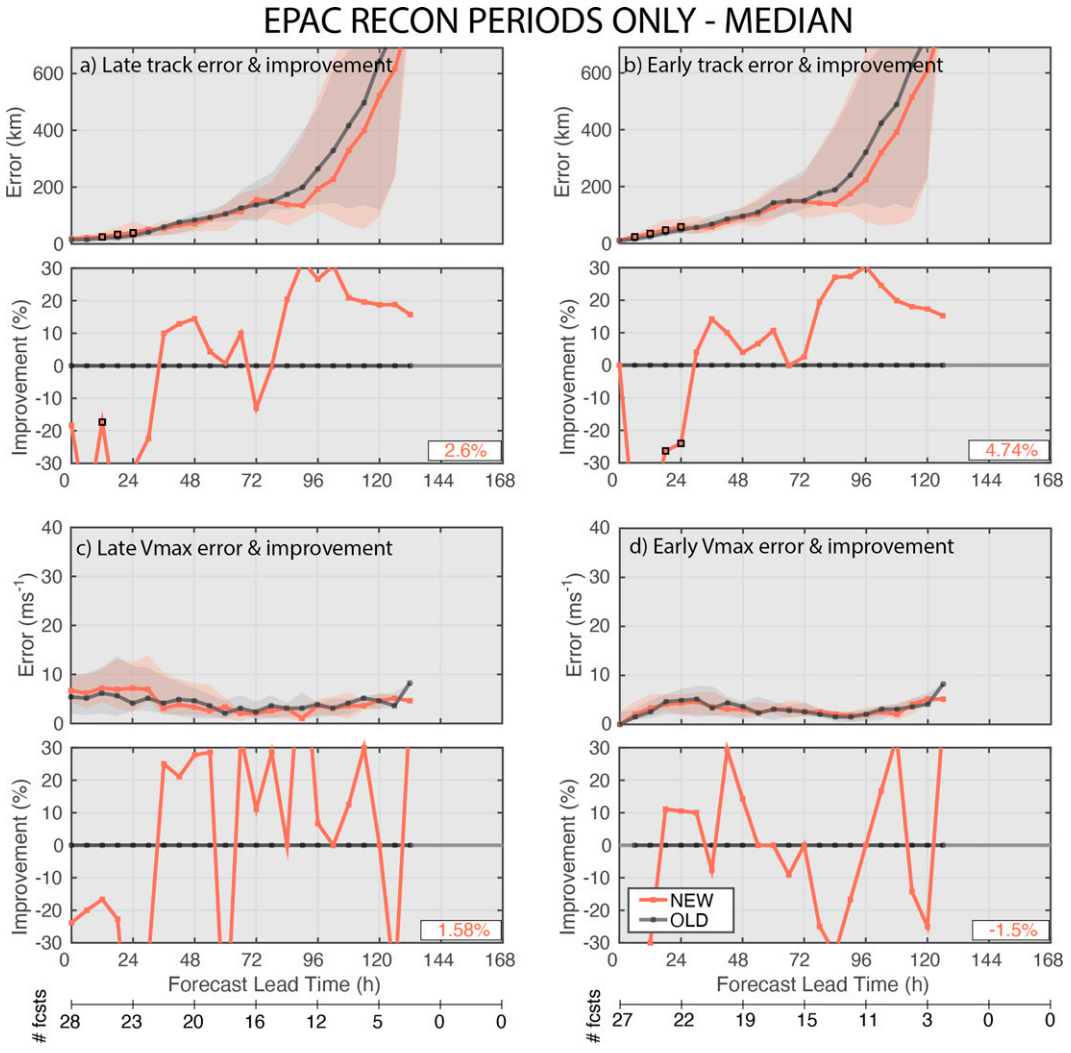

FIG. 15. As in Fig. 5, but for EPAC reconnaissance periods.

consistent with the finding that the added data in NEW significantly alters the track but not the intensity.

4. EPAC results

We now examine the impact of assimilating additional reconnaissance data in NEW on various aspects of EPAC TC forecasts. Given the very small sample size of EPAC cases with reconnaissance in this study, the approach to analysis in this section is a bit different than above. In particular, we examine the cases with reconnaissance (Figs. 14–15) and the full sample (Figs. 16–18) but do not perform additional stratifications.

a. EPAC reconnaissance periods

Reconnaissance missions sample EPAC tropical systems much less frequently than NATL systems. With a few exceptions for research, missions only occur in the event of a hurricane threat to the west coast of Mexico or for a tropical storm or hurricane threat to Hawaii (NOAA 2021). During the periods examined here, generally short periods of reconnaissance occurred for only 4 storms, as opposed to 13 storms with some long-duration periods in the NATL sample (Fig. 1b). This limits the EPAC sample size in this section to about 10% of that in the NATL, thus limiting the robustness of this dataset.

Generally speaking, assimilating additional reconnaissance data does not directly improve forecasts of EPAC TCs in this sample. In both the early and late models (Fig. 14), EPAC short-term track forecasts significantly degrade with the additional data, but they improve at later lead times. The late NEW intensity forecast is also worse than that of OLD at short lead times (Fig. 14c), but similar to the NATL results, the early model intensity in NEW (Fig. 14d) improves upon the late model. The additional data in NEW results in slightly larger negative bias in the late intensity forecast (Fig. 14e), but the interpolator adequately removes that bias (Fig. 14f). Results are similar in terms of median error (Fig. 15) with the exception that NEW performs better than OLD at extended lead times, though the sample size is tiny.

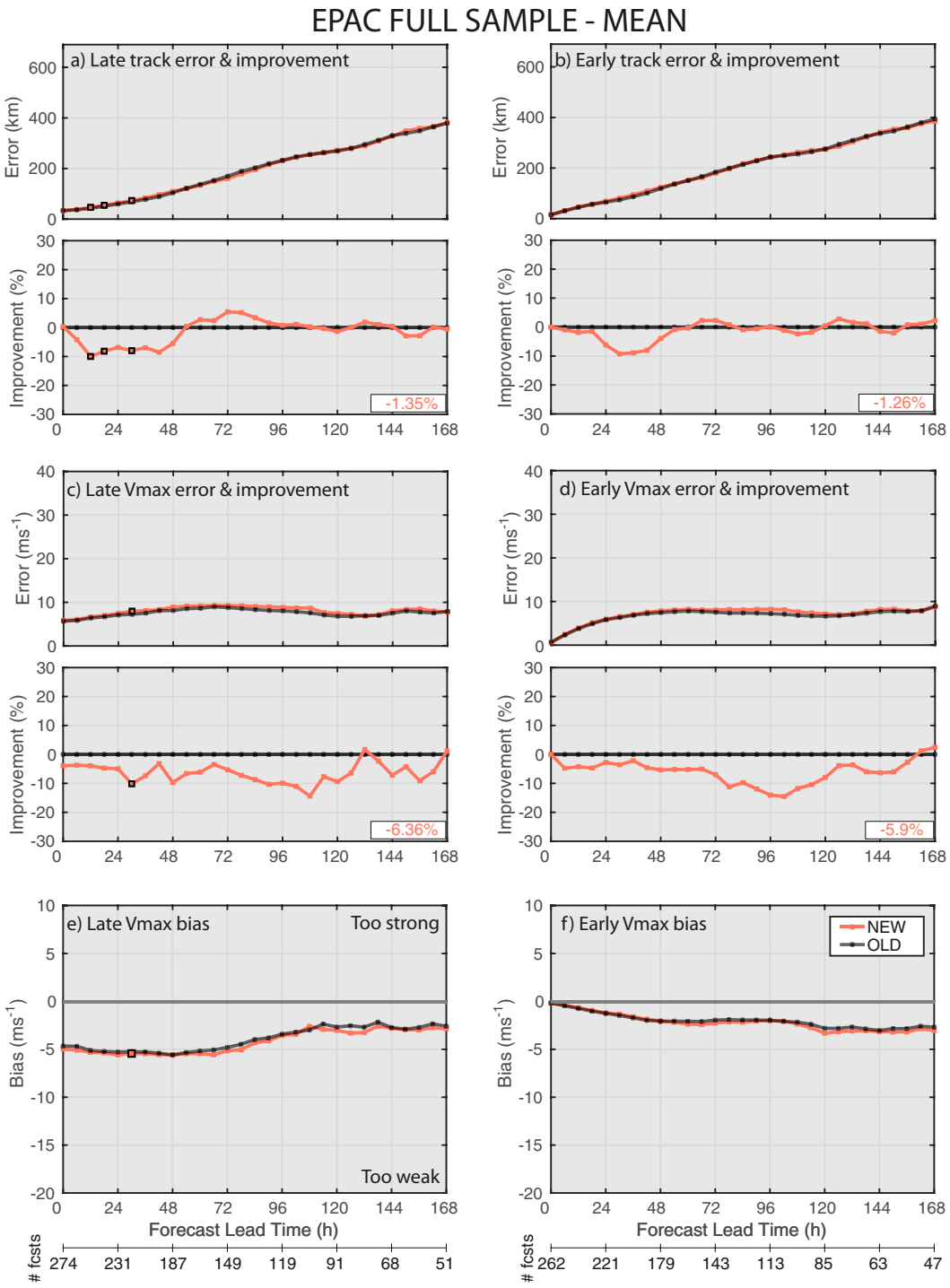

FIG. 16. As in Fig. 4, but for the entire EPAC sample.

The reason for the discrepancy between the EPAC and NATL results is unclear. A few possibilities include fundamentally different storm characteristics, intermittent EPAC sampling, different approaches to reconnaissance in the EPAC (e.g., generally less inner-core sampling), or simply sampling error. Perhaps little should be read into the EPAC results here since the sample size is small, and half the sample comes from one storm—Olivia (2018).

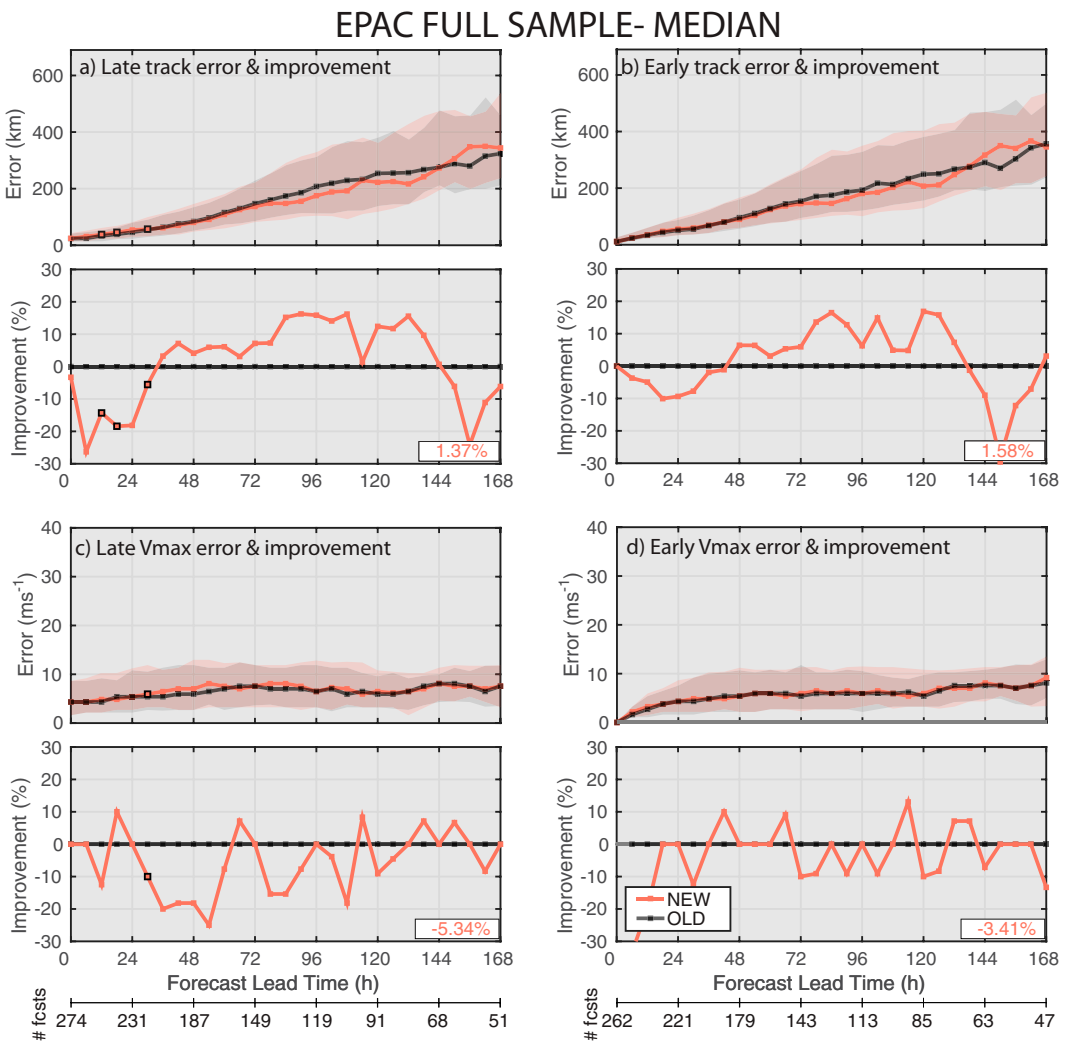

FIG. 17. As in Fig. 5, but for the entire EPAC sample.

b. *Entire EPAC sample*

Remote influences strongly drive EPAC results here. As mentioned in section 4a, only 10% of the cycles in Figs. 16–17 occurred during reconnaissance periods directly sampling EPAC storms. Thus, the remainder generally occurred during periods of NATL reconnaissance (cf. Tables 2–4 EPAC column three to NATL column four).

Unlike the NATL sample, adding reconnaissance data has mixed impacts in the full EPAC sample and particularly varies according to whether verifying the mean or median error. The short-term track degrades significantly in NEW compared with OLD (Fig. 16a), though the actual MAE difference at these lead times is only about 5 km. The interpolator does end up removing some, but not all, of this degradation in NEW (Fig. 16b). After 48 h, OLD and NEW have very similar mean track errors. Though the median short-term track errors also increase in NEW (Figs. 17a,b), they improve upon those in OLD after 36 h. Similar to track, additional reconnaissance (primarily in NATL) impacts mean and median EPAC intensity errors differently. The mean intensity errors in NEW are substantially larger than those in OLD, though only one lead time yields statistically significant differences in Fig. 16c. The median intensity errors in Figs. 17c,d differ less between the two experiments, particularly in the early forecast.

As with the NATL track errors at extended lead times, histograms shed light on the disparity between mean and median performance of NEW relative to OLD. Figure 18 shows the distribution of absolute track and intensity errors at 96 h. For the track, NEW has a much larger peak frequency than OLD in the bin centered on 100 km and generally lower frequencies in the higher error bins. The exception is at 500 km, where NEW has a number of outlier errors. Thus, the median error in NEW outperforms OLD considerably, but the means are nearly identical. Likewise, NEW and OLD have a similar peak around 5 m s^{-1} Vmax error, but NEW has a few more outliers in the 30–35 m s^{-1} bins. These outliers degrade the

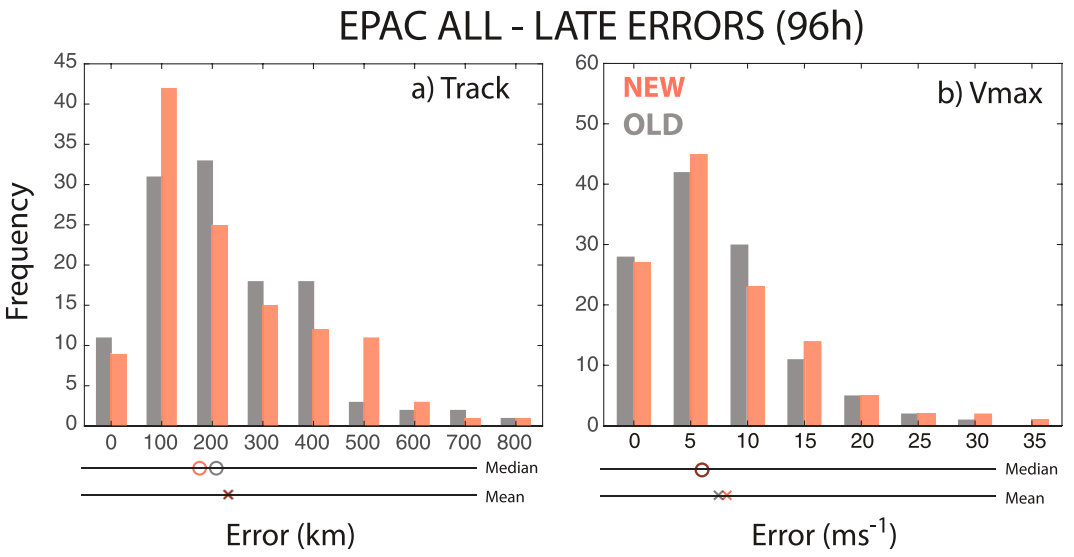

FIG. 18. Histograms of EPAC (a) track and (b) intensity absolute errors in the full sample of OLD and NEW at 96 h. The mean and median of the distributions are indicated below each histogram.

MAE relative to OLD, but they do not meaningfully impact the median.

5. Discussion and conclusions

This study has found that assimilating additional reconnaissance data (flight-level HDOBs and dropwindsonde wind data near the TC center) has profound impacts on subsequent track forecasts in the NCEP GFSv16 model. For NATL TCs, statistically significant improvements in track extend through 4–5 days during periods of reconnaissance. Even upon expanding to the entire NATL sample to include storms and periods without reconnaissance, significant improvement remains at some lead times. Further assessment, discussed below, suggests that greater improvements might also be expected at the longer lead times (e.g., days 6–7).

Aside from examining the impact of reconnaissance data, this study also explores the importance of comprehensively assessing data impact. For example, model or data assimilation changes can affect the so-called "early" and "late" versions of the forecast very differently. In this instance, adding data degraded the short-term late intensity forecasts, but that degradation was a result of negative bias. The postprocessing that occurs to generate the early forecast removes such short-term biases such that a forecaster never sees them in real time (though some forecasters do use information from the late model output, the early output has a much more immediate and direct impact on the actual operational forecast). In another example, this study demonstrates the importance of exploring different ways to describe the error statistics. In several instances discussed below, the impacts of the additional data strongly differ depending on whether one examines the mean or median errors.

The first example of incongruence in impact assessed from mean and median errors comes in the day 6–7 lead times. The MAE in Fig. 4 gives the impression that the reconnaissance impact begins to wane in the longer range forecast. However, the median errors shown in Fig. 5 suggest that might not be true. The forecast at long lead times has a fairly small sample size and is more susceptible to outliers (e.g., Fig. 6). As such, the 15%–30% median improvement seen on those days in Fig. 5 might more accurately represent what could be expected in a larger sample.

Another example of the importance of assessing median errors comes from the EPAC results, particularly for intensity. The MAE in Figs. 16c and 16d leaves the distinct impression that adding NATL reconnaissance observations degrades EPAC track and intensity forecasts. However, the median errors (Figs. 17c,d) tell a different story, suggesting that even this larger sample is susceptible to outliers. Taking into account the available evidence, it appears that the overall impact on EPAC TCs in this study is mixed.

The results here are congruent with a large body of previous work that has demonstrated the positive impacts of reconnaissance data. Among studies that have examined HDOB impact, WZ16 probably provides the most comparable results since they used a large sample focused exclusively on reconnaissance periods and assessed similar observation types. They showed that the combination of HDOB and dropwindsondes improved track forecasts by up to 15% in a research data assimilation and forecast system. Nevertheless, there are several major differences between WZ16 and the current study. First, they used a regional system where the added data could only affect the analysis within the vicinity of each storm. Further, they assessed the impact of all dropswindondes and HDOBs in their analysis domain.

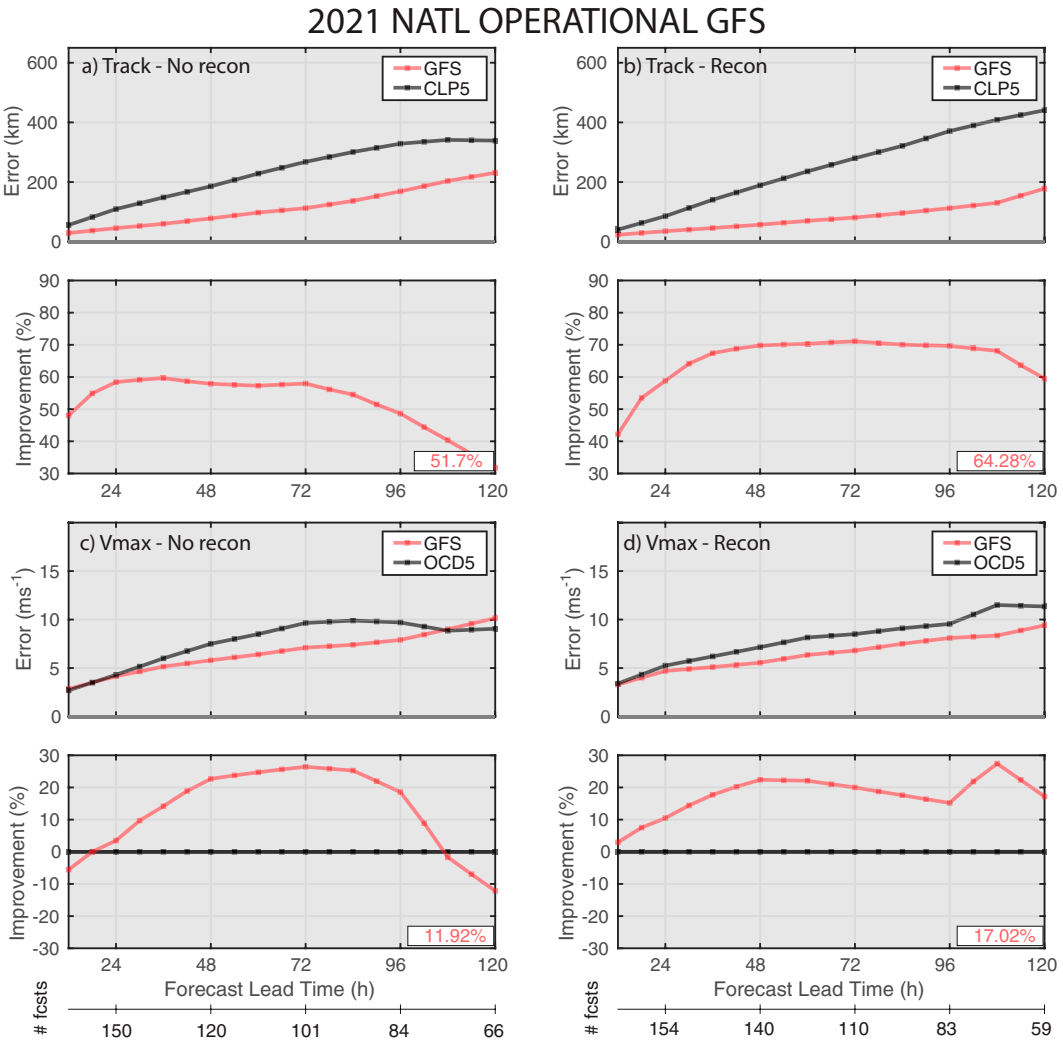

FIG. 19. Performance of the operational GFSv16 during the 2021 NATL hurricane season during storms and periods (left) without reconnaissance and (right) with reconnaissance. Improvement is evaluated against the CLP5 baseline for track and OCD5 baseline for intensity.

Combined with the previous published work on dropwindsondes, the results of this study suggest that the net impact of all currently used reconnaissance in GFSv16 (i.e., HDOBs and all dropwindsondes) might exceed what WZ16 found. Though such an assessment does not exist for GFSv16, here we compare the track and intensity performance of GFSv16 during the 2021 NATL hurricane season during and outside of reconnaissance periods. In particular, Fig. 19, respectively, shows GFSv16 track and intensity skill with respect to the CLP5 and OCD5 climatological and persistence models. The left side of Fig. 19 shows verification for storms and periods of no reconnaissance, while the right side shows similar verification for storms and periods with reconnaissance. This assessment carries the caveat that the samples with and without reconnaissance cover different cases so that other impacts such as nearness to the U.S. observation network cannot be accounted for. Nevertheless, 2021 GFSv16 track forecasts were much more skillful than CLP5 when reconnaissance was available. Most notable is the improvement at the extended lead times, where the difference in skill approaches 30%. Meanwhile, intensity skill relative to OCD5 in the reconnaissance period generally does not improve except at perhaps the extended lead times.

Improved use of available reconnaissance data is likely necessary to meet current operational goals. For example, the Hurricane Forecast Improvement Project (HFIP) recently set forth a number of ambitious 10-yr goals (Marks et al. 2019) in response to the Weather Research and Forecasting Innovation Act of 2017. Goals related to numerical weather prediction included:

- Reduce numerical forecast guidance errors by 50% from 2017;
- Produce 7-day forecast guidance similar to 2017 5-day forecast guidance;

- Improve guidance on preformation disturbances, including genesis timing and track and intensity forecasts, by 20% from 2017.

Though the goals are ambitious, the results here imply that they are achievable with improved use of available data along with data assimilation and model improvements. While adding HDOB data and some additional dropwindsonde observations has certainly improved GFSv16 performance with TCs, there remains a great deal of reconnaissance data that the model still does not use.

In an example directly relevant to this study, the use of dropwindsondes themselves at NCEP needs improvement. Though NOAA reconnaissance missions transmit high-resolution dropwindsonde data in real time, global DA at NCEP uses information only in the body of WMO TEMP DROP messages, which only contain the dropwindsonde release point to the nearest tenth of a degree. Thus, an entire dropwindsonde is assimilated in a column with a somewhat inaccurate initial location, and not considering the dropwindsonde drift. Since inner-core dropwindsondes in hurricanes sometimes travel to the opposite side of the storm (e.g., Aberson 2008), assimilating such wind data at the wrong location produces extreme, erroneous analysis increments. The current way NCEP processes dropwindsonde data prevents inner-core dropwindsondes in hurricanes from being used (Aberson 2008; Aberson et al. 2017), even in this study. Further, the TEMP DROP messages only contain data at mandatory and significant levels, which ongoing work suggests could limit their impact (K. Sellwood 2022, personal communication).

Beyond enhanced use of dropwindsonde data, other improvements related to reconnaissance data usage can likely improve GFSv16 forecasts further. In particular, neither operationally transmitted tail Doppler radar (TDR) data from the NOAA reconnaissance planes (WP-3Ds and G-IV) nor SFMR data are assimilated in the NCEP global DA system. Both of these observation types have improved forecasts in the operational HWRF and other research data assimilation systems (e.g., Weng and Zhang 2012; Aberson et al. 2015). Further, a number of research instruments on the NOAA reconnaissance planes provide valuable observations that are not currently transmitted to NCEP in real time. Among this data are wind profiles from the Imaging Wind and Rain Airborne Profiler (Guimond et al. 2014) and from a Doppler wind lidar (Bucci 2020; Bucci et al. 2018; Zhang et al. 2018), which provide valuable data that fill in current observational gaps.

Other research has shown the tremendous value of assimilating data from unmanned aerial systems (UAS) as well. Wick et al. (2020) showed that data transmitted from the Global Hawk UAS had fairly large positive impacts on both the HWRF and GFS models. Of note, Global Hawk dropwindsonde data improved TC track forecasts by up to 15% in GFSv15. Key benefits of the Global Hawk include its high altitude (~60 000 ft), and its long endurance (~24 h) ensures that nearly the entire tropical North Atlantic falls within its sampling range. Meanwhile, Aksoy et al. (2022) demonstrated benefits of assimilating data from a small UAS launched during WP-3D missions into Hurricane Maria (2017). Ongoing unpublished work has shown that these data also benefit the operational HWRF model. The above evidence suggests that UASs will likely play a key role in future TC forecast improvement.

For the many cases that reconnaissance or UASs cannot cover, improving assimilation from other data sources can also significantly improve TC forecasts. Groundbreaking research has shown that better use of oceanic data sources can reduce forecast errors during rapid intensification (Domingues et al. 2021; Le Hénaff et al. 2021), while other work has shown that TC forecasts benefit from improved used of satellite-derived winds (e.g., Velden et al. 2017; Sawada et al. 2019; Lim et al. 2019; Lewis et al. 2020; Li et al. 2020) and satellite radiance data (Zhang et al. 2016; Minamide and Zhang 2017, 2018; Zhang et al. 2021).

In conclusion, this study has demonstrated the tremendous potential for further improving TC forecasts. In addition to the benefits shown here, many further improvements await with improved use of these and other reconnaissance observations. This demonstrates the need of not only investing in data assimilation improvements, but also enhancements to observational systems in order to reach next-generation hurricane forecasting goals.

Acknowledgments. The authors thank Frank Marks, Sim Aberson, Matt Rigney, and three anonymous reviewers for constructive comments that contributed a great deal to this manuscript.

Data availability statement. The ATCF files used to compile the error statistics for NEW and OLD in this manuscript can be retrieved from https://www.emc.ncep.noaa.gov/gc_wmb/wd20xw/HDOB/ATCF/. The CLP5 and OCD5 forecasts are available in the a-decks from http://hurricanes.ral.ucar.edu/repository/. The final B-decks (i.e., best tracks) used for verification are available from NHC and can be found at https://www.nhc.noaa.gov/data/hurdat. The GRaphics for OS(S)Es and Other modeling applications on TCs (GROOT) verification package developed by the third author (Ditchek) and funded by the Quantitative Observing System Assessment Program (QOSAP) and the FY18 Hurricane Supplemental (NOAA Award ID NA 19OAR0220188) was used with some modifications to generate graphics for this publication. It can be found at https://github.com/sditchek/GROOT.

REFERENCES

Aberson, S. D., 2008: Large forecast degradations due to synoptic surveillance during the 2004 and 2005 hurricane seasons. *Mon. Wea. Rev.*, **136**, 3138–3150, https://doi.org/10.1175/2007MWR2192.1.

——, 2010: 10 years of hurricane synoptic surveillance (1997–2006). *Mon. Wea. Rev.*, **138**, 1536–1549, https://doi.org/10.1175/2009MWR3090.1.

——, A. Aksoy, K. J. Sellwood, T. Vukicevic, and X. Zhang, 2015: Assimilation of high-resolution tropical cyclone observations with an ensemble Kalman filter using HEDAS: Evaluation of

2008–11 HWRF forecasts. *Mon. Wea. Rev.*, **143**, 511–523, https://doi.org/10.1175/MWR-D-14-00138.1.

——, K. J. Sellwood, and P. A. Leighton, 2017: Calculating dropwindsonde location and time from TEMP-DROP messages for accurate assimilation and analysis. *J. Atmos. Oceanic Technol.*, **34**, 1673–1678, https://doi.org/10.1175/JTECH-D-17-0023.1.

Aksoy, A., J. C. Cione, B. Dahl, and P. D. Reasor, 2022: Tropical cyclone data assimilation with Coyote uncrewed aircraft system observations, very-frequent cycling, and a new online quality control technique. *Mon. Wea. Rev.*, **150**, 797–820, https://doi.org/10.1175/MWR-D-21-0124.1.

Alaka, G. J., Jr., X. Zhang, S. G. Gopalakrishnan, S. B. Goldenberg, and F. D. Marks, 2017: Performance of basin-scale HWRF tropical cyclone track forecasts. *Wea. Forecasting*, **32**, 1253–1271, https://doi.org/10.1175/WAF-D-16-0150.1.

Bucci, L. R., 2020: Assessment of the utility of Doppler wind lidar for tropical cyclone analysis and forecasting. Ph.D. dissertation, University of Miami, 132 pp.

——, C. O'Handley, G. D. Emmitt, J. A. Zhang, K. Ryan, and R. Atlas, 2018: Validation of an airborne Doppler wind lidar in tropical cyclones. *Sensors*, **18**, 4288, https://doi.org/10.3390/s18124288.

Burpee, R. W., D. G. Marks, and R. T. Merrill, 1984: An assessment of omega dropwindsonde data in track forecasts of Hurricane Debby (1982). *Bull. Amer. Meteor. Soc.*, **65**, 1050–1058, https://doi.org/10.1175/1520-0477(1984)065<1050:AAOODD>2.0.CO;2.

——, J. L. Franklin, S. J. Lord, R. E. Tuleya, and S. D. Aberson, 1996: The impact of omega dropwindsondes on operational hurricane track forecast models. *Bull. Amer. Meteor. Soc.*, **77**, 925–934, https://doi.org/10.1175/1520-0477(1996)077<0925:TIOODO>2.0.CO;2.

Cangialosi, J. P., 2022. National Hurricane Center Forecast Verification Report: 2021 hurricane season. NOAA/NHC, 76 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf.

Domingues, R., and Coauthors, 2021: Ocean conditions and the intensification of three major Atlantic hurricanes in 2017. *Mon. Wea. Rev.*, **149**, 1265–1286, https://doi.org/10.1175/MWR-D-20-0100.1.

Farrar, M., 2021a: Upgrade NCEP Global Forecast Systems (GFS) to v16. NWS Service Change Notice 21-20. NOUS41 KWBC 181950 AAC, 13 pp., https://www.weather.gov/media/notification/pdf2/scn21-20_gfsv16.0_aac.pdf.

——, 2021b: Upgrade NCEP Global Forecast System to v16.1.1. NWS Service Change Notice 21-52. NOUS41 KWBC 171715, 2 pp., https://www.weather.gov/media/notification/pdf2/scn21-52_gfsv16.1.1.pdf.

Guimond, S. R., L. Tian, G. M. Heymsfield, and S. J. Frasier, 2014: Wind retrieval algorithms for the IWRAP and HIWRAP airborne Doppler radars with applications to hurricanes. *J. Atmos. Oceanic Technol.*, **31**, 1189–1215, https://doi.org/10.1175/JTECH-D-13-00140.1.

Kleist, D. T., 2011: Assimilation of tropical cyclone advisory minimum sea level pressure in the NCEP global data assimilation system. *Wea. Forecasting*, **26**, 1085–1091, https://doi.org/10.1175/WAF-D-11-00045.1.

——, and Coauthors, 2021: NCEP operational global data assimilation upgrades: From versions 15 through 16. *Special Symp. on Global and Mesoscale Models: Updates and Center Overviews WAF Symp. General Session*, online, Amer. Meteor. Soc., 12.3, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/378554.

——, and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

Le Hénaff, M., and Coauthors, 2021: The role of the Gulf of Mexico ocean conditions in the intensification of Hurricane Michael (2018). *J. Geophys. Res. Oceans*, **126**, e2020JC016969, https://doi.org/10.1029/2020JC016969.

Lei, L., and J. S. Whitaker, 2016: A four-dimensional incremental analysis update for the ensemble Kalman filter. *Mon. Wea. Rev.*, **144**, 2605–2621, https://doi.org/10.1175/MWR-D-15-0246.1.

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

Lewis, W. E., C. S. Velden, and D. Stettner, 2020: Strategies for assimilating high-density atmospheric motion vectors into a regional tropical cyclone forecast model (HWRF). *Atmosphere*, **11**, 673, https://doi.org/10.3390/atmos11060673.

Li, J., J. Li, C. Velden, P. Wang, T. J. Schmit, J. Sippel, 2020: Impact of rapid-scan-based dynamical information from GOES-16 on HWRF hurricane forecasts. *J. Geophys. Res. Atmos.*, **125**, e2019JD031647, https://doi.org/10.1029/2019JD031647.

Lim, A. H. N., J. A. Jung, S. E. Nebuda, J. M. Daniels, W. Bresky, M. Tong, and V. Tallapragada, 2019: Tropical cyclone forecasts impact assessment from the assimilation of hourly visible, shortwave, and clear-air water vapor atmospheric motion vectors in HWRF. *Wea. Forecasting*, **34**, 177–198, https://doi.org/10.1175/WAF-D-18-0072.1.

Marchok, T. P., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., P1.13, https://ams.confex.com/ams/25HURR/techprogram/paper_37628.htm.

Marks, F., N. Kurkowski, M. DeMaria, and M. Brennan, 2019: Hurricane Forecast Improvement Program Five-Year Plan: 2019–2024: Proposed Framework for Addressing Section 104 of the Weather Research Forecasting Innovation Act of 2017. NOAA, 86 pp., https://hfip.org/sites/default/files/documents/hfip-strategic-plan-20190625-final.pdf.

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

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

Minamide, M., and F. Zhang, 2017: Adaptive observation error inflation for assimilating all-sky satellite radiance. *Mon. Wea. Rev.*, **145**, 1063–1081, https://doi.org/10.1175/MWR-D-16-0257.1.

——, and ——, 2018: Assimilation of all-sky infrared radiances from *Himawari-8* and impacts of moisture and hydrometer initialization on convection-permitting tropical cyclone prediction. *Mon. Wea. Rev.*, **146**, 3241–3258, https://doi.org/10.1175/MWR-D-17-0367.1.

NOAA, 2021: National hurricane operations plan. OFCM Doc. FCM-P12-2021, NOAA, 185 pp., https://www.icams-portal.gov/resources/ofcm/nhop/2021_full_nhop_change_2.pdf.

Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419, https://doi.org/10.1175/2008WAF2222128.1.

Sawada, M., Z. Ma, A. Mehra, V. Tallapragada, R. Oyama and K. Shimoji, 2019: Impacts of assimilating high-resolution atmospheric motion vectors derived from *Himawari-8* on tropical cyclone forecast in HWRF. *Mon. Wea. Rev.*, **147**, 3721–3740, https://doi.org/10.1175/MWR-D-18-0261.1.

Tong, M., and Coauthors, 2018: Impact of assimilating aircraft reconnaissance observations on tropical cyclone initialization and prediction using operational HWRF and GSI ensemble–variational hybrid data assimilation. *Mon. Wea. Rev.*, **146**, 4155–4177, https://doi.org/10.1175/MWR-D-17-0380.1.

Trahan, S., and L. Sparling, 2012: An analysis of NCEP tropical cyclones vitals and potential effects on forecasting models. *Wea. Forecasting*, **27**, 744–756, https://doi.org/10.1175/WAF-D-11-00063.1.

Velden, C., W. E. Lewis, W. Bresky, D. Stettner, J. Daniels, and S. Wanzong, 2017: Assimilation of high-resolution satellite-derived atmospheric motion vectors: Impact on HWRF forecasts of tropical cyclone track and intensity. *Mon. Wea. Rev.*, **145**, 1107–1125, https://doi.org/10.1175/MWR-D-16-0229.1.

Weng, Y., and F. Zhang, 2012: Assimilating airborne Doppler radar observations with an ensemble Kalman filter for convection-permitting hurricane initialization and prediction: Katrina (2005). *Mon. Wea. Rev.*, **140**, 841–859, https://doi.org/10.1175/2011MWR3602.1.

——, and ——, 2016: Advances in convection-permitting tropical cyclone analysis and prediction through EnKF assimilation of reconnaissance aircraft observations. *J. Meteor. Soc. Japan*, **94**, 345–358, https://doi.org/10.2151/jmsj.2016-018.

Wick, G. A. and Coauthors, 2020: NOAA's Sensing Hazards with Operational Unmanned Technology (SHOUT) experiment observations and forecast impacts. *Bull. Amer. Meteor. Soc.*, **101**, E968–E987, https://doi.org/10.1175/BAMS-D-18-0257.1.

Yang, F., V. S. Tallapragada, D. T. Kleist, A. Chalwa, J. Wang, R. Treadon, and J. Whitaker, 2021: On the development and evaluation of NWS global forecast systems version 16. *Special Symp. on Global and Mesoscale Models: Updates and Center Overviews WAF Symp. General Session*, online, Amer. Meteor. Soc., 12.2, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/378135.

Zawislak, J., and Coauthors, 2021: Accomplishments of NOAA's Airborne Hurricane Field Program and a Broader Future Approach to Forecast Improvement. *Bull. Amer. Meteor. Soc.*, **103**, E311–E338, https://doi.org/10.1175/BAMS-D-20-0174.1.

Zhang, F., Y. Weng, J. F. Gamache, and F. D. Marks, 2011: Performance of convection-permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner core airborne Doppler radar observations. *Geophys. Res. Lett.*, **38**, L15810, https://doi.org/10.1029/2011GL048469.

——, M. Minamide, and E. E. Clothiaux, 2016: Potential impacts of assimilating all-sky infrared satellite radiances from GOES-R on convection-permitting analysis and prediction of tropical cyclones. *Geophys. Res. Lett.*, **43**, 2954–2963, https://doi.org/10.1002/2016GL068468.

Zhang, J. A., R. Atlas, G. D. Emmitt, L. Bucci, and K. Ryan, 2018: Airborne Doppler wind lidar observations of the tropical cyclone boundary layer. *Remote Sens.*, **10**, 825, https://doi.org/10.3390/rs10060825.

Zhang, Y., and Coauthors, 2021: Ensemble-based assimilation of satellite all-sky microwave radiances improves intensity and rainfall predictions for Hurricane Harvey (2017). *Geophys. Res. Lett.*, **48**, e2021GL096410, https://doi.org/10.1029/2021GL096410.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the next generation global prediction system. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Zhu, Y., J. C. Derber, R. J. Purser, B. A. Ballish, and J. Whiting, 2015: Variational correction of aircraft temperature bias in the NCEP's GSI analysis system. *Mon. Wea. Rev.*, **143**, 3774–3803, https://doi.org/10.1175/MWR-D-14-00235.1.