A Systematic Assessment of the Overall Dropsonde Impact during the 2017–20 Hurricane Seasons Using the Basin-Scale HWRF

SARAH D. DITCHEK,[a,b] JASON A. SIPPEL,[b] GHASSAN J. ALAKA JR.,[b] STANLEY B. GOLDENBERG,[b] AND LIDIA CUCURULL[b]

[a] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[b] *NOAA/AOML/Hurricane Research Division, Miami, Florida*

(Manuscript received 6 June 2022, in final form 21 February 2023, accepted 6 March 2023)

ABSTRACT: This study comprehensively assesses the overall impact of dropsondes on tropical cyclone (TC) forecasts. We compare two experiments to quantify dropsonde impact: one that assimilated and another that denied dropsonde observations. These experiments used a basin-scale, multistorm configuration of the Hurricane Weather Research and Forecasting Model (HWRF) and covered active North Atlantic basin periods during the 2017–20 hurricane seasons. The importance of a sufficiently large sample size as well as thoroughly understanding the error distribution by stratifying results are highlighted by this work. Overall, dropsondes directly improved forecasts during sampled periods and indirectly impacted forecasts during unsampled periods. Benefits for forecasts of track, intensity, and outer wind radii were more pronounced during sampled periods. The forecast improvements of outer wind radii were most notable given the impact that TC size has on TC-hazards forecasts. Additionally, robustly observing the inner- and near-core region was necessary for hurricane-force wind radii forecasts. Yet, these benefits were heavily dependent on the data assimilation (DA) system quality. More specifically, dropsondes only improved forecasts when the analysis used mesoscale error covariance derived from a cycled HWRF ensemble, suggesting that it is a vital DA component. Further, while forecast improvements were found regardless of initial classification and in steady-state TCs, TCs undergoing an intensity change had diminished benefits. The diminished benefits during intensity change probably reflect continued DA deficiencies. Thus, improving DA system quality and observing system limitations would likely enhance dropsonde impacts.

SIGNIFICANCE STATEMENT: This study uses a regional hurricane model to conduct the most comprehensive assessment of the impact of dropsondes on tropical cyclone (TC) forecasts to date. The main finding is that dropsondes can improve many aspects of TC forecasts if their data are assimilated with sufficiently advanced assimilation techniques. Particularly notable is the impact of dropsondes on TC outer-wind-radii forecasts, since improving those forecasts leads to more effective TC-hazard forecasts.

KEYWORDS: Hurricanes/typhoons; Dropsondes; Forecast verification/skill; Data assimilation; Model evaluation/performance; Numerical weather prediction/forecasting

1. Introduction

Over the past several decades, numerous peer-reviewed studies have quantified the impact of tropical cyclone (TC) airborne reconnaissance data on numerical weather prediction (NWP) model forecasts. These types of studies are needed since simply adding additional data to forecast models, even high-quality data, does not ensure improved forecasts. Perhaps most valuable for understanding the impact of reconnaissance data are studies that used large, multiyear samples, as these are less subject to inherent variability among TCs or even years (e.g., Aberson 2010; Weng and Zhang 2016; Tong et al. 2018; Zawislak et al. 2022; Sippel et al. 2022). While these studies varied in the type of reconnaissance data assessed (i.e., individual types of reconnaissance data and overall), they all found that airborne reconnaissance data generally benefit TC forecasts, and these benefits increase with more advanced data assimilation (DA) and modeling systems.

Despite the documented benefits of reconnaissance data, the above studies leave some important questions unanswered. For example, only two of the above studies assessed how reconnaissance impacts vary by TC intensity [Tong et al. (2018, hereafter T18) and Sippel et al. (2022)]. In particular, T18 found degraded reconnaissance impact in strong TCs in the 2013 version of the National Centers for Environmental Prediction (NCEP) Hurricane Weather Research and Forecasting Model (HWRF). Though the shortcomings that caused that degradation seem to have since been addressed (e.g., Zawislak et al. 2022), no subsequent assessment has thoroughly reexamined the intensity-dependent impact of reconnaissance data on TC forecasts using HWRF. Further, while most of the studies listed above assessed the impacts of reconnaissance during sampled periods, only Sippel et al. (2022) compared that direct impact to the effect of reconnaissance on TCs during unsampled periods.

From the suite of currently operationally available reconnaissance data,[1] it is particularly useful to understand the

Corresponding author: Sarah D. Ditchek, sarah.d.ditchek@noaa.gov

[1] This includes high-density observations (HDOBs) from flight level (e.g., temperature, specific humidity, and wind) and from the stepped frequency microwave radiometer (SFMR), radial velocity from the tail Doppler radar (TDR), and pressure, temperature, relative humidity, and horizontal winds from global positioning satellite (GPS) dropsondes.

DOI: 10.1175/WAF-D-22-0102.1

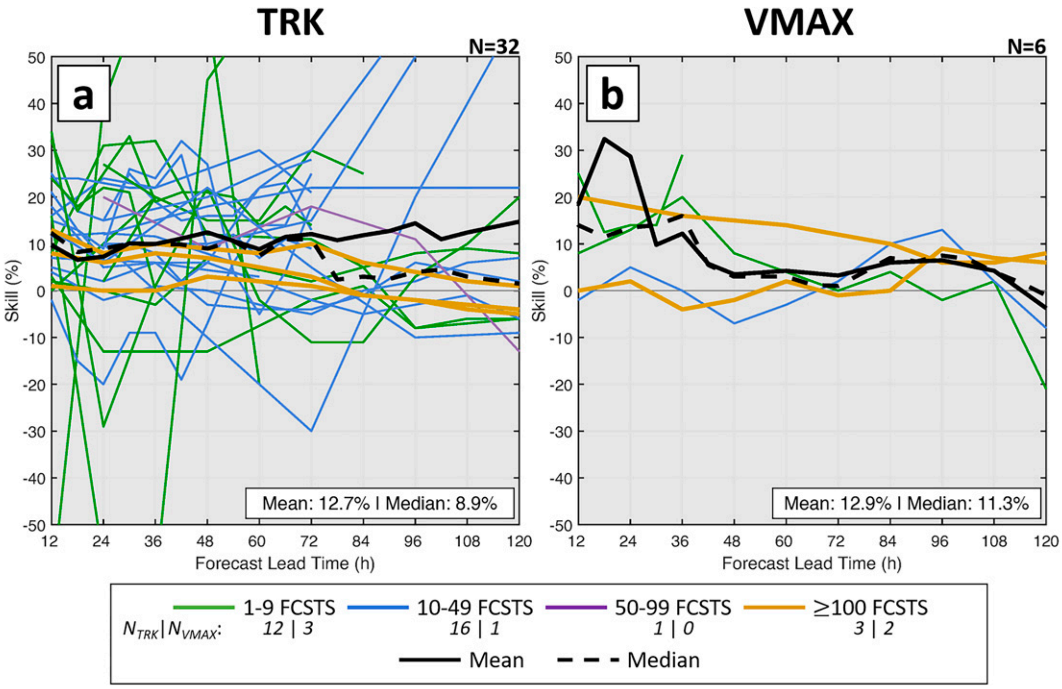


FIG. 1. Skill from 32 OSEs in the literature that quantified the overall impact of dropsondes on TC forecasts of (a) track (TRK) and (b) maximum sustained 10-m wind speed (VMAX). The *x* axis is forecast lead time. The *y* axis is relative skill with respect to experiments without assimilated dropsonde observations. The number of OSEs in each graphic is given in the graphic title as *N*. Note that the *y*-axis maximum in (b) is capped at 50 despite one OSE having skill values that were far outside the *y*-axis limits. The colored lines indicate the number of individual forecasts (FCSTS) in each experiment, with the number of studies for both variables with 1–9, 10–49, 50–99, or ≥100 individual forecasts given in the legend. Note that those OSEs with 100 or more individual forecasts (yellow) are given a thicker line, for contrast. The mean (black solid line) and median (black dashed line) across all OSEs are also shown, with their averages given in the bottom-right corner. Results from the following papers are included: Franklin and DeMaria (1992), Burpee et al. (1996), Shi et al. (1996), Aberson and Franklin (1999), Aberson (2002), Aberson and Etherton (2006), Wu et al. (2007), Pu et al. (2008), Yamaguchi et al. (2009), Harnisch and Weissmann (2010), Aberson (2010, 2011), Chou et al. (2011), Wu et al. (2012), and Majumdar et al. (2013).

impacts of global positioning system (GPS) dropsondes on TC forecasts. Dropsondes are expendable and come with a substantial cost (around $800 each; M. Brennan 2022, personal communication), and since at least 20–30 dropsondes are launched per flight, they add considerably to the cost of a typical airborne mission. Yet, this added cost could be justified if dropsondes improve TC forecasts and subsequent warnings and help reduce evacuation footprints, economic impact, and loss of life. Thus, numerous studies have assessed the overall[2] impact of dropsonde on TC forecasts.

Previous research has generally found that overall, dropsondes improve TC track and intensity forecasts in both regional and global models. Yet, a major shortcoming of those studies is that large sample sizes have rarely been used. This makes the systematic assessment of their impacts difficult due to year-to-year, TC-to-TC, and even forecast-to-forecast variability. To demonstrate this, Fig. 1 both summarizes the overall impact of dropsondes from previous observing system experiments (OSEs; i.e., data-denial experiments with real data) in peer-reviewed literature[3] and highlights this sample-size drawback. Dropsonde impacts have generally been positive for track (Fig. 1a), though results varied substantially between individual experiments. The very small sample sizes used by nearly all OSEs included in Fig. 1 perhaps contributed to this variability. Note that only three track-impact assessments used a sample of 100 or more cases (Fig. 1a; thick yellow lines; Aberson 2010, 2011). Previous work assessing dropsonde impacts on TC maximum sustained 10-m wind

[2] Studies have also assessed the impact of dropsondes from specific aircraft (e.g., NASA's high-altitude Global Hawk; Christophersen et al. 2017, 2018; Kren et al. 2018; Wick et al. 2020) and the impact of dropsondes that target specific regions near the TC or in the synoptic environment (e.g., Aberson 2003; Yamaguchi et al. 2009; Majumdar 2016; Torn 2020, 2021). Generally, improvement was found in both global and regional models due to dropsonde assimilation.

[3] These studies were from 15 papers (see Fig. 1 caption) covering the years 1992–2022 that documented the overall impact of dropsondes on TC forecasts. Note that they had large variability in the type of model used (e.g., global or regional), the DA method applied (e.g., three- or four-dimensional variational or ensemble–variational), the sample size, and the source of the dropsondes (e.g., field campaigns or specific aircraft).

speed is even more limited (Fig. 1b), with only two large-sample OSEs conducted (Fig. 1b; thick yellow lines; Aberson 2010). Not shown is that only two OSEs quantified the impact of dropsondes on minimum sea level pressure, though they only assessed the impact through 36 h, and neither used a large sample. Finally, no study has quantitatively assessed the impact of dropsondes on forecasts of significant wind radii associated with TCs. This is a particularly important omission since the size of a tropical cyclone correlates strongly with its damage potential (Powell and Reinhold 2007, e.g.,) and is important for TC-hazards forecasts and warnings.

To address the shortcomings found in the literature for both reconnaissance and dropsonde data-impact studies, this study conducts the most comprehensive assessment of the overall impact of dropsondes on NWP-model TC track, intensity, and significant-wind-radii forecasts to date. In doing so, it also represents the most comprehensive assessment of the impact of any airborne observing system on TC forecasts to date. Here, impact was quantified during active periods in the North Atlantic basin (NATL; including the North Atlantic Ocean, the Gulf of Mexico, and the Caribbean Sea) within the 2017–20 hurricane seasons using the 2020 version of the basin-scale, multistorm configuration of HWRF. The full sample was first stratified to assess data impact during sampled and unsampled periods and any interannual variability present. Further stratifications allowed for comparison with previous studies such as T18 and facilitated a qualitative assessment of the DA system. Among these are stratifications by initial Saffir–Simpson-scale classification (Simpson and Saffir 1974), ongoing intensity change, and the covariance choice used for inner-core DA.

Though this study does not directly examine sensitivity to the details of the DA system, many of the results here strongly suggest that DA limitations modulate the benefits of dropsondes. This warrants a brief review of pertinent inner core DA studies. Specifically, using appropriate mesoscale error covariance is crucial for improving tropical cyclone forecasts. Zhang et al. (2009) first showed this with an experimental research system that used a cycled ensemble Kalman filter (EnKF). More recently, Lu et al. (2017a, hereafter L17) showed that assimilating inner-core data with error covariance from the NCEP Global Data Assimilation System (GDAS) did not improve forecasts of Hurricane Sandy (2012). On the other hand, assimilating the same data with mesoscale error covariance derived from an EnKF-cycled HWRF ensemble (hereafter, HWRF-cycled covariance) generally did improve the forecasts. Further, T18 used GDAS covariance in their much larger reconnaissance impact assessment and found major problems when assimilating inner-core reconnaissance data in HWRF. As described above, the reconnaissance data significantly degraded forecasts of hurricanes, in part due to suboptimal error covariance for inner-core DA. In some cases, the degradation extended through the entire forecast. As described in detail in section 2a, the operational HWRF has advanced considerably since T18, most notably by using HWRF-cycled covariance for particular TCs. As such, this study presents an excellent opportunity to compare with the results from L17 and T18.

The rest of this paper is organized as follows: section 2 describes the data and methods used, section 3 focuses on the impact of dropsondes on TC forecasts in the NATL, section 4 briefly discusses the impact of dropsondes on TC forecasts in the eastern North Pacific basin (EPAC), and section 5 provides a summary and recommendations for future work.

2. Data and methods

This section provides a description of the model and DA system (section 2a), the dropsonde observing systems (section 2b), and the experiment setup and scope (section 2c). Also detailed are the verification methods used (section 2d) including stratifications of the full sample taken [section 2d(1)] and metrics used in analysis [section 2d(2)]. The consistency metric introduced in Ditchek et al. (2023) will be used for analysis and is briefly detailed in section 2d(2).

a. *Model and DA system*

HWRF is a triply nested, nonhydrostatic, coupled atmosphere–ocean NWP model capable of producing accurate, high-resolution forecasts every 6 h for TCs. The outermost domain (i.e., the parent domain; D01) captures the evolution of the environment around the TC, while the telescopic nests (D02 and D03) follow TCs to better simulate their evolution. The 2020 version of HWRF (H220) used respective resolutions of 13.5, 4.5, and 1.5 km for D01, D02, and D03 along with 75 vertical levels and a model top of 10 hPa. HWRF includes a three-dimensional ensemble–variational (3DEnVar) hybrid DA system based on a Gridpoint Statistical Interpolation (GSI) analysis system to produce an analysis from which the forecast is initialized. Physics schemes utilized include the Ferrier–Aligo cloud microphysics scheme, a scale-aware Simplified Arakawa–Schubert (SAS) cumulus parameterization, the Rapid Radiative Transfer Model for GCMs (RRTMG) radiation scheme for both longwave and shortwave, the Noah land surface model (LSM), the GFDL model surface layer scheme, and the GFS hybrid eddy-diffusivity mass-flux (EDMF) planetary boundary layer parameterization. The horizontal localization correlation length varies by domain, being 300 and 150 km for D02 and D03, respectively. The vertical localization correlation length is 0.5 ln(p) for both domains, where p is in centibars (cbar). For more details on the model configuration and the DA system, please see (T18 and Biswas et al. 2018).

The "basin-scale" HWRF (HWRF-B; Zhang et al. 2016; Alaka et al. 2017, 2019, 2020, 2022) used in this study is an experimental, parallel, multistorm version of HWRF that was developed by NOAA's Hurricane Research Division (HRD) of the Atlantic Oceanographic and Meteorological Laboratory (AOML) in collaboration with the NOAA NCEP Environmental Modeling Center (EMC) and the Developmental Testbed Center (DTC). HWRF-B is identical to HWRF except in two key ways (Alaka et al. 2022):

- HWRF-B contains a large, static parent domain that covers nearly the entire area of responsibility of NOAA's National Hurricane Center (NHC; i.e., the North Atlantic and eastern North Pacific hurricane basins), while HWRF has a

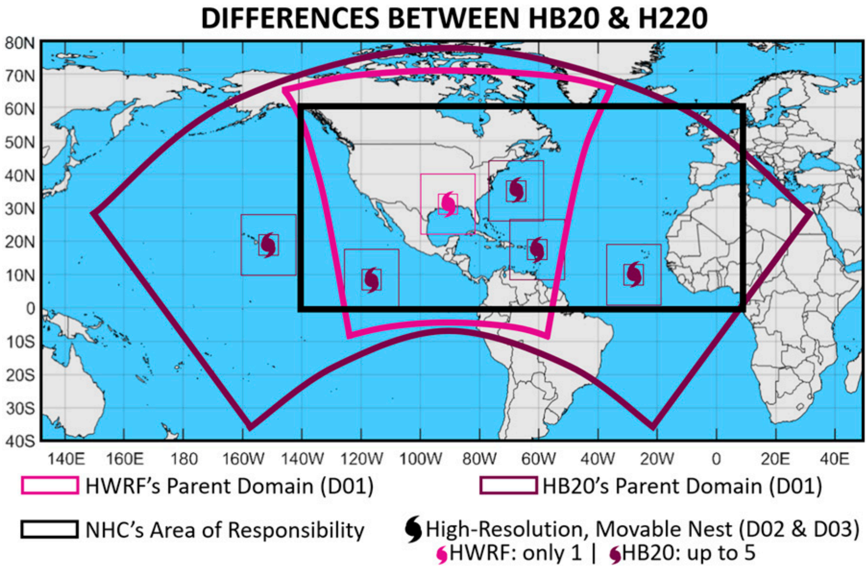

DIFFERENCES BETWEEN HB20 & H220
80N
70N
60N
50N
40N
30N
20N
10N
0
10S
20S
30S
40S
140E
160E
180
160W
140W
120W
100W
80W
60W
40W
20W
0
20E
40E
HWRF's Parent Domain (D01)
HB20's Parent Domain (D01)
NHC's Area of Responsibility
High-Resolution, Movable Nest (D02 & D03)
HWRF: only 1 | HB20: up to 5

FIG. 2. Visualization of the domain and high-resolution, movable nest differences between HB20 (dark purple) and H220 (pink). The black box shows NHC's area of responsibility including the NATL and EPAC.

smaller parent domain that relocates from one forecast to the next.

- HWRF-B features TC-following, telescopic nests for multiple TCs per model integration, while HWRF uses only one set of telescopic nests for one TC per model integration.

These two differences enable HWRF-B to better resolve TC-TC interactions, synoptic-scale features, TC genesis, and other features (Alaka et al. 2022). For details on how the performance of HWRF-B compares to HWRF, see Alaka et al. (2017, 2020, 2022).

The 2020 version of HWRF-B (HB20) used in this study is identical to H220 except in the two ways described above. In HB20, TC-following nests are configured for up to five TCs within the same outermost domain (Fig. 2). They are triggered by the presence of disturbances or TCs in the tropical cyclone vitals database (TCVitals[4]). For each domain, HB20 uses a different initialization procedure as related to DA. The parent domain (D01) has no DA—it is initialized by directly interpolating the global analysis onto D01. On the other hand, the TC-following nests (D02 and D03) do use DA. For D02, the ensemble covariance in GSI comes from the previous cycle's 80-member NCEP's Global DA System (GDAS) ensemble 6-h forecasts. For D03, as with H220, the source of the ensemble covariance depends on the TC priority.

For "low-priority" TCs (i.e., TCs that are expected to remain weak, have minimal impacts and/or do not have reconnaissance data), the previous cycle's 80-member NCEP GDAS ensemble 6-h forecasts provides the D03 flow-dependent covariance. This is largely the same configuration described in T18. While this covariance choice for D03 is suboptimal, it has performed reasonably well for weaker TCs.

For "high-priority" TCs (i.e., TCs that are forecast to intensify, be high impact, and/or have reconnaissance data), a more advanced procedure is used. In particular, the D03 flow-dependent covariance is calculated from a 40-member D02 ensemble native to HB20 whose perturbations are updated each cycle with an EnKF. This configuration, which is similar to that of L17, benefits intensity forecasts since GDAS covariance alone can lead to short-term negative intensity biases that are especially problematic for stronger TCs (L17 and T18). As discussed in section 1, a major intent of using HWRF-cycled covariance is to benefit situations with inner-core data.

Computational constraints in operations do not permit HWRF-cycled covariance to run for more than one TC at a time. Thus, if more than one TC has ongoing reconnaissance, the TC with the greatest overall threat uses HWRF-cycled covariance, while the others use GDAS covariance. To facilitate comparison with HWRF, HB20 uses a similar concept as in operations. Thus, some of the TCs with reconnaissance in this study also use GDAS covariance for DA.

b. *Dropsonde observing systems*

This study assesses the impacts of dropsondes released into TCs between 2017 and 2020 from four different types of aircraft: the U.S. Air Force Reserve's low–midaltitude WC-130J (C-130), NOAA's low–midaltitude WP-3D (P-3), NOAA's high-altitude Gulfstream IV-SP (G-IV), and NASA's high-altitude Global Hawk (GH).[5] The resulting

[4] TCVitals contains operational estimates of a TC's position, intensity, significant wind radii, and motion (Trahan and Sparling 2012).

[5] Only two GH flights occurred during the time period of this study: one during Harvey (2017) and another during Lidia (2017).

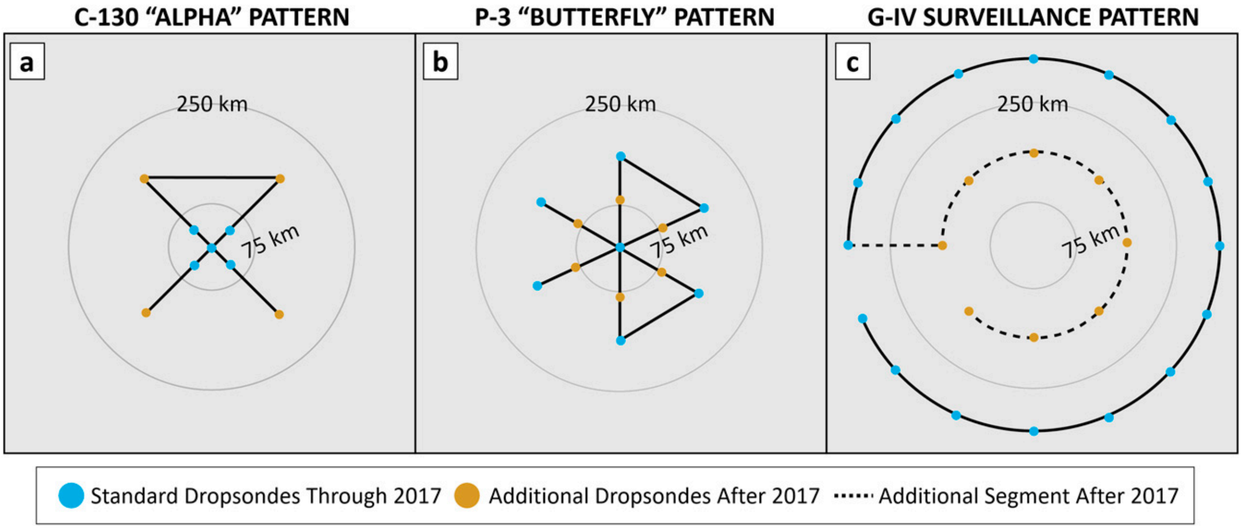

C-130 "ALPHA" PATTERN
P-3 "BUTTERFLY" PATTERN
G-IV SURVEILLANCE PATTERN
a
b
c
250 km
75 km
250 km
75 km
250 km
75 km
● Standard Dropsondes Through 2017
● Additional Dropsondes After 2017
······ Additional Segment After 2017

FIG. 3. Representative examples of the (a) C-130 "alpha" pattern, (b) P-3 "butterfly" pattern, and (c) G-IV circumnavigation-ring surveillance pattern, with dropsondes included through 2017 in blue, additional dropsondes added after 2017 in gold, and the additional G-IV circumnavigation added after 2017 as the dashed black line. Note that (c) does not include those G-IV dropsondes released in the large-scale environment.

atmospheric profiles of quality-controlled pressure, temperature, relative humidity, and horizontal winds were transmitted in TEMP DROP messages from the aircraft (NOAA 2020).

To account for the horizontal advection of dropsondes with height (i.e., dropsonde drift), both H220 and its parallel HB20 assimilate dropsonde data at horizontal locations to the nearest hundredth of a degree based on an algorithm originally developed at HRD (Aberson et al. 2017). That algorithm uses information from the 62626 section of the TEMP DROP message as well as wind data from the main body of the TEMP DROP message for the location estimates, which are typically accurate to within 0.5 km. This step is necessary since dropsonde locations used elsewhere at NCEP (e.g., in the operational GFS) are derived from the main body of TEMP DROP messages, which contains only the initial release point to the nearest tenth of a degree. Since dropsondes can sometimes travel azimuthal distances exceeding 180° in the eyewall of a hurricane, assimilating dropsonde observations as a profile can result in unphysical analysis increments (Aberson 2008). Note that this postprocessed data only includes the same mandatory and significant levels from the original TEMP DROP message.

Four substantial changes in dropsonde TC sampling occurred during the 2017–20 period. Beginning experimentally in 2017 and implemented operationally in 2018, dropsondes were released at the end points of C-130 alpha-pattern formations (around 150–200 km from the TC center). Also beginning in 2018, midpoint dropsondes released at a radius of around 80 km became fairly routine on P-3 missions. Additionally, in 2017 the G-IV only conducted one circumnavigation around TCs at around 330 km from the TC center. Beginning in 2018, the G-IV began to conduct an additional circumnavigation at around 165 km when possible (Sippel 2020; Sippel et al. 2021). Figure 3 gives representative examples of the C-130, P-3, and G-IV flight patterns with an indication of observing system changes after 2017. Finally, beginning experimentally in 2017 and implemented operationally in 2019, ensemble sensitivity metrics described in Torn (2020, 2021) have guided environmental sampling conducted by the G-IV. Because of these substantial changes in dropsonde usage over the period of assessment, this study also investigates the interannual differences in dropsonde impact.

c. *Experiment setup and scope*

Two experiments together quantify the impact of dropsondes on TC forecasts: 1) the "All Dropsondes" experiment (ALL; i.e., all available dropsonde observations were assimilated if they did not fail quality control checks in GSI[6]) and 2) the "No Dropondes" experiment (NO; i.e., dropsonde observations were not assimilated). Both experiments otherwise assimilated all conventional, reconnaissance, and satellite data assimilated into H220 (Ditchek et al. 2022).

To maximize the number of 120-h forecasts included in the assessment, each experiment ran for eight forecast blocks[7] covering active NATL periods within the 2017–20 hurricane seasons (Table 1). Briefly summarizing the table, experiments

[6] In H220 and HB20 GSI rejects observations when the difference between the first guess and observation (i.e., innovation) is larger than a predefined multiple of the observation error. Observation errors for dropsondes are inflated linearly proportional to the magnitude of the innovation so that very few observations actually get rejected. Those rejected are typically winds in the eyewall, but even there an overwhelming majority are assimilated.

[7] A forecast block is defined as a continuous set of forecasts in which at least one TC is active.

TABLE 1. Years run, forecast block (FB), start date, end date, cycles, the number of TCs with assimilated dropsonde observations (DRP), the total number of TCs (NTC), and TCs broken down into NATL an EPAC. Asterisks (*) indicate TCs with assimilated dropsonde observations. Plus signs (+) indicate "high-priority" TCs that used HWRF-cycled covariance. A summary row is included underneath the table.

Year	FB	Start date	End date	Cycles	DRP/NTC	NATL	EPAC
2017	1	0600 UTC 16 Aug 2017	0000 UTC 29 Sep 2017	176	8/20	08L (Gert), 09L (Harvey)$^{*+}$ 10L (Ten)$^{+}$, 11L (Irma)$^{*+}$ 12L (Jose)$^{*+}$, 13L (Katia)* 14L (Lee), 15L (Maria)$^{*+}$ Invests (92L*, 99L)	13E (Kenneth), 14E (Lidia)* 15E (Otis), 16E (Max) 17E (Norma)*, 18E (Pilar) Invests (94E, 95E, 98E, 99E) —
2018	2	0600 UTC 30 Aug 2018	0600 UTC 15 Sep 2018	65	5/10	06L (Florence)$^{*+}$, 07L (Gordon)$^{*+}$ 08L (Helene), 09L (Isaac)$^{*+}$ 10L (Joyce), Invest (95L)	15E (Miriam), 16E (Norman)$^{*+}$ 17E (Olivia)*, 18E (Paul) —
	3	1200 UTC 23 Sep 2018	1200 UTC 13 Oct 2018	81	4/11	11L (Eleven), 12L (Kirk)* 13L (Leslie)$^{+}$, 14L (Michael)$^{*+}$ 15L (Nadine), Invests (92L, 98L)	20E (Rosa)*, 21E (Sergio)$^{*+}$ 22E (Tara) Invest (97E)
2019	4	1200 UTC 18 Aug 2019	0000 UTC 9 Sep 2019	87	5/14	04L (Chantal), 05L (Dorian)$^{*+}$ 06L (Erin)*, 07L (Fernand)$^{*+}$ 08L (Gabrielle) Invests (90L, 92L*, 94L, 97L)	10E (Ivo)$^{*+}$, 11E (Juliette) 12E (Akoni) Invests (96E, 97E) —
	5	0000 UTC 14 Sep 2019	1200 UTC 2 Oct 2019	75	6/11	09L (Humberto)$^{*+}$, 10L (Jerry)$^{*+}$ 12L (Karen)$^{*+}$, 13L (Lorenzo)$^{*+}$ Invest (96L)	13E (Kiko), 14E (Mario)* 15E (Lorena)*, 16E (Narda) Invests (93E, 95E)
2020	6	1800 UTC 24 Jul 2020	0600 UTC 5 Aug 2020	47	4/6	07L (Gonzalo)*, 08L (Hanna)$^{*+}$ 09L (Isaias)$^{*+}$, 10L (Ten) Invest (94L)	08E (Douglas)$^{*+}$ — —
	7	1800 UTC 16 Aug 2020	0600 UTC 29 Aug 2020	51	4/9	11L (Josephine), 13L (Laura)$^{*+}$ 14L (Marco)* —	10E (Ten), 11E (Fausto) 12E (Genevieve)*, 13E (Hernan) 14E (Iselle), Invests (98E)*
	8	0600 UTC 11 Sep 2020	0000 UTC 24 Sep 2020	52	5/11	17L (Paulette)$^{*+}$, 18L (Rene) 19L (Sally)$^{*+}$, 20L (Teddy)$^{*+}$ 21L (Vicky)*, 22L (Beta)* 23L (Wilfred), 24L (Alpha)	16E (Karina), 17E (Lowell) Invest (91E) — —
					Summary		
4 years	8 FBs	159 days		634 cycles	41/92 with dropsondes	29/53 with dropsondes	12/39 with dropsondes

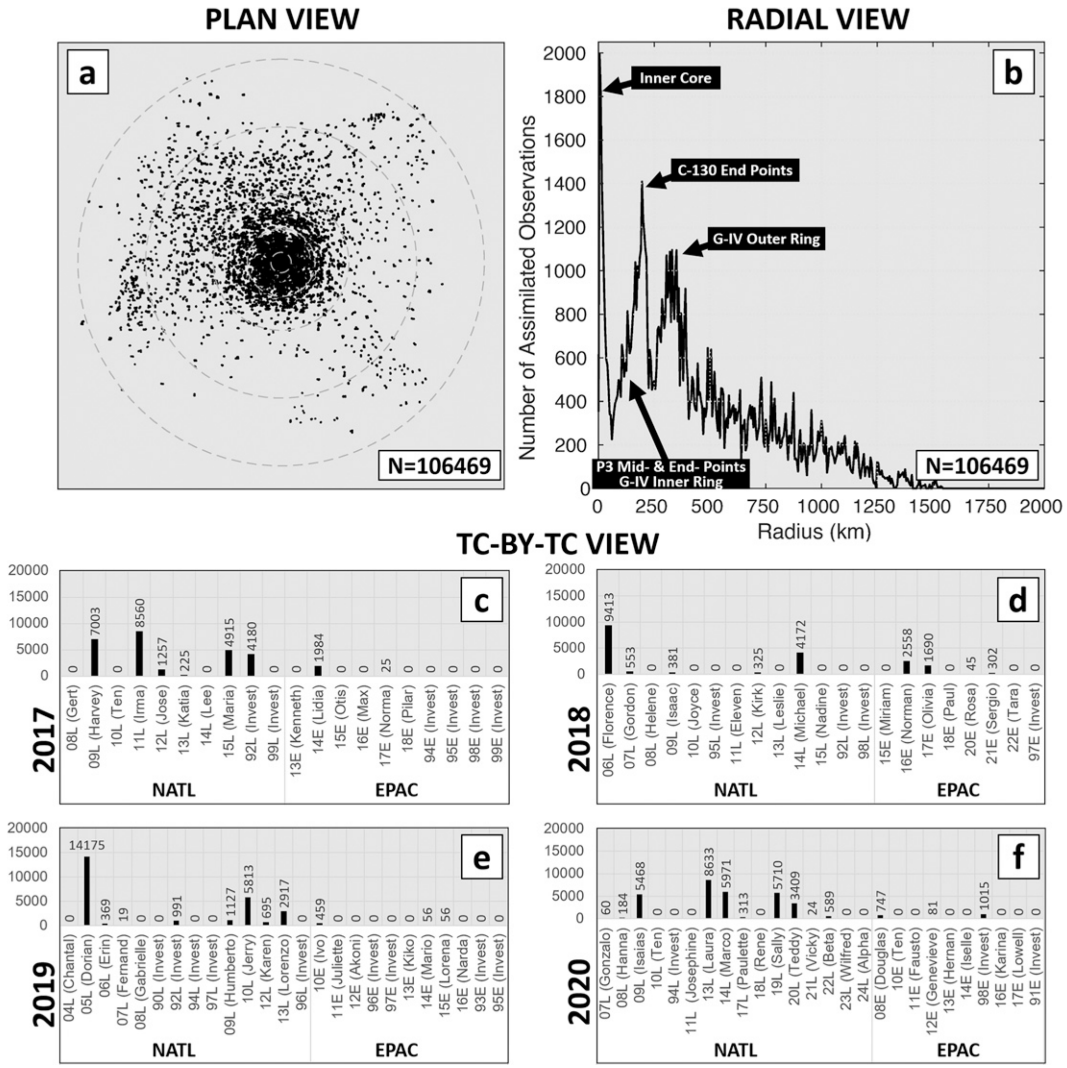


FIG. 4. The number of individually assimilated dropsonde temperature observations in each TC's D02 for the full sample in a (a) TC-relative plan view, (b) TC-relative radial view, and TC-by-TC view for (c) 2017, (d) 2018, (e) 2019, and (f) 2020.

cover about 159 days during the 4-year period and include 634 cycles, resulting in 2139 individual forecasts across 92 TCs. Of those 92 TCs, 41 had dropsonde observations that were assimilated in ALL. Since HB20 allows TC-following nests for up to five TCs per cycle, assimilating dropsonde observations in one TC affects all concurrent and subsequent TC forecasts (Alaka et al. 2017, 2020). For the number of assimilated dropsonde observations by TC, see Figs. 4c–f.

Figure 4 depicts the number of individually assimilated dropsonde temperature[8] observations in each TC's D02 between 2017 and 2020 in a: 1) TC-relative plan view, 2) TC-relative radial view, and 3) TC-by-TC view. In Fig. 4a, an axisymmetric distribution of assimilated temperature observations is observed within 500 km. Between 500 and 1000 km, there is a northward bias from sampling the synoptic environment ahead of the TC track, mainly by the G-IV. Note that the same dropsonde could be assimilated into multiple TCs with overlapping D02 domains. Consequently, data points in Fig. 4 do not correspond 1 to 1 with actual dropsonde data, especially at radii > 1000 km. For example, temperature observations assimilated in the inner core of Harvey (2017) were also assimilated in nearby Invest 92L's D02 at large radii (not shown). In Fig. 4b a frequency peak at radii < 75 km corresponds to the inner-core dropsondes at the center and around the radius of maximum wind. Radially outward from there, an increase in the between the first and second peaks corresponds with dropsondes at the P-3 mid- and end points as well as the G-IV inner ring. A second peak around 200 km corresponds with the most frequent location of C-130 end points. Finally, a third peak just outside the 300-km radius corresponds with

[8] Only temperature is shown for simplicity, though temperature, humidity, and winds were all assimilated.

TABLE 2. The nine "overall" and "TC-by-TC" metrics that are calculated and used to analyze results.

Category	Metric	Description
Overall	MAE	The mean of the absolute-valued difference between an experiment's forecast and the best track at the forecast verifying time
	MDAE	The median of the absolute-valued difference between an experiment's forecast and the best track at the forecast verifying time
	MAE skill	On average how much better or worse an experiment performed over a baseline experiment using the MAE
	MDAE skill	The median of how much better or worse an experiment performed over a baseline experiment using the MDAE
	FSP	The percent of forecasts where an experiment outperformed a baseline experiment
TC-by-TC	MAE skill	The MAE skill for individual TCs
	MDAE skill	The MDAE skill for individual TCs
	FSP	The FSP for individual TCs
	Sample size	The sample size at each lead time for individual TCs

G-IV outer-ring circumnavigation dropsondes. The distribution then tapers off to zero around the 1500-km radius.

d. Verification

This study evaluates the performance of each experiment separately for each basin by verifying forecasts against the NHC "best track" (Rappaport et al. 2009) available from NHC following the standard NHC forecast verification procedures. Forecasts are included if at both the initialization time and the forecast verifying time the system was classified by the final best track as a tropical or subtropical cyclone (Cangialosi 2022). While NHC only verifies individual forecasts that initialize east of 140°W in their EPAC samples, this study includes all TCs initialized between 140°W and 180°. Results presented in this paper are for raw output from the GFDL vortex tracker (Marchok 2002, 2021) only, without any additional postprocessing (e.g., interpolation to produce "early" model forecasts; Cangialosi 2022). Note that current TC-verification techniques do not account for uncertainties in position, intensity, and significant wind radii that are present in both the best track (Torn and Snyder 2012; Landsea and Franklin 2013) and HWRF tracker output (Zhang et al. 2021). Finally, NHC Forecast Verification procedures reduced the 2139 individual forecasts in the sample to 1032 verifiable NATL forecasts and 535 verifiable EPAC forecasts.

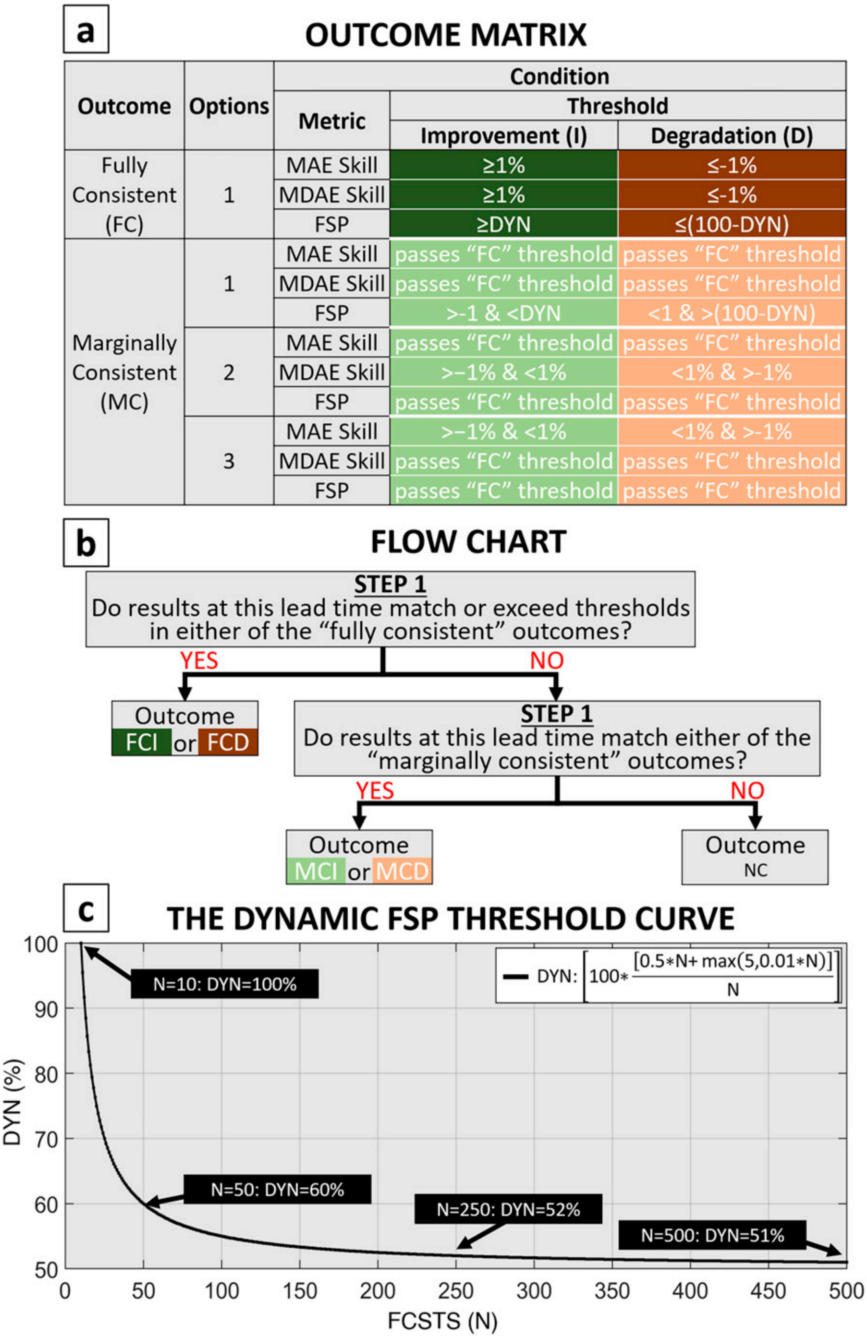
Variables assessed include track and two measures of TC intensity [maximum sustained 10-m wind speed (VMAX); minimum sea level pressure (PMIN)], as well as the three significant surface wind radii reported by NHC [34-kt wind radii (R34), 50-kt wind radii (R50), and 64-kt wind radii (R64); 1 kt ≈ 0.51 m s^{-1})]. Note that samples of R34, R50, and R64 include all quadrants. Finally, this paper presents results from homogeneous samples. To be included in the homogeneous sample: 1) for a given forecast, all experiments have to satisfy the standard NHC forecast verification procedures and 2) a nonzero numeric value has to exist for a given variable in all experiments. This second condition only impacts R34, R50, and R64 samples.

1) STRATIFICATIONS

For each variable, this study stratifies the full sample by a number of different factors. The first stratification is by data availability into sampled and unsampled periods. Those forecasts with dropsonde observations assimilated in D02 at 0 h (i.e., sampled periods) are identified as forecasts with direct dropsonde impact (hereafter, OBS). The remaining forecasts (i.e., forecasts without dropsonde observations assimilated in D02 at 0 h) are categorized as forecasts with indirect dropsonde impact (hereafter, NOOBS). This indirect dropsonde impact can occur in two ways: 1) dropsondes assimilated in the nests for one TC can impact forecasts of other unsampled TCs in the parent domain and 2) dropsondes assimilated in previous cycles can impact future cycles. Both OBS and NOOBS are further stratified by year to demonstrate how yearly differences impacted the full sample.

Since previous studies have found that the choice of D03 covariance strongly modulates reconnaissance data improvements (e.g., L17 and T18), this study also stratifies OBS by D03 covariance used: HWRF-cycled (i.e., high-priority TCs) or GDAS (i.e., low-priority TCs). For brevity, this manuscript only shows stratification by covariance for the OBS subset, as NOOBS results did not meaningfully depend on covariance choice (not shown). Hereafter, the subset of OBS that utilized the HWRF-cycled covariance will be referred to as OBS-HCOV, and the subset that used GDAS will be referred to as OBS-GCOV. Note that stratification by covariance yields a fairly small sample size (<100 cases) for OBS-GCOV—since this study focused on active NATL periods, around 80% of forecasts with assimilated dropsondes at in D02 at 0 h used HWRF-cycled covariance. Further, stratifying OBS-HCOV by year or classification also yields undesirably small samples, so neither OBS-GCOV or OBS-HCOV are further stratified. Finally, while it is possible that the covariance stratification is subject to sampling bias, the results obtained are qualitatively similar to previously published work (discussed below).

To explore when sampling TCs had the most impact and to enable comparison to previous work (i.e., T18), OBS is also



Outcome	Options	Condition		
		Metric	Threshold	
			Improvement (I)	Degradation (D)
Fully Consistent (FC)	1	MAE Skill	≥1%	≤-1%
		MDAE Skill	≥1%	≤-1%
		FSP	≥DYN	≤(100-DYN)
Marginally Consistent (MC)	1	MAE Skill	passes “FC” threshold	passes “FC” threshold
		MDAE Skill	passes “FC” threshold	passes “FC” threshold
		FSP	>-1 & <DYN	<1 & >(100-DYN)
	2	MAE Skill	passes “FC” threshold	passes “FC” threshold
		MDAE Skill	>–1% & <1%	<1% & >-1%
		FSP	passes “FC” threshold	passes “FC” threshold
	3	MAE Skill	>–1% & <1%	<1% & >-1%
		MDAE Skill	passes “FC” threshold	passes “FC” threshold
		FSP	passes “FC” threshold	passes “FC” threshold

b
FLOW CHART
STEP 1
Do results at this lead time match or exceed thresholds in either of the “fully consistent” outcomes?
YES
NO
Outcome
FCI or FCD
STEP 1
Do results at this lead time match either of the “marginally consistent” outcomes?
YES
NO
Outcome
MCI or MCD
Outcome
NC
c
THE DYNAMIC FSP THRESHOLD CURVE
DYN: [100∗ [0.5∗N+ max(5,0.01∗N)] / N]
N=10: DYN=100%
N=50: DYN=60%
N=250: DYN=52%
N=500: DYN=51%
DYN (%)
100
90
80
70
60
50
0
50
100
150
200
250
300
350
400
450
500
FCSTS (N)

FIG. 5. This figure is from Ditchek et al. (2023) and includes (a) the outcome matrix of consistency metric conditions and associated thresholds (where “DYN” represents the dynamic FSP threshold) for whether the sample has fully consistent (one option) or marginally consistent (three possible options) improvement or degradation at each forecast lead time in one experiment relative to a baseline experiment, (b) a flowchart for determining the outcome for a forecast lead time, and (c) values of the dynamic FSP threshold as a function of the number of individual forecasts (FCSTS) based on the equation in the legend. Since the dynamic FSP threshold will always be 51% for $N \geq 500$, the x axis terminates at $N = 500$. Also, since FSP cannot exceed 100%, the y axis terminates at 100%. For more details on the consistency metric, see Ditchek et al. (2023).

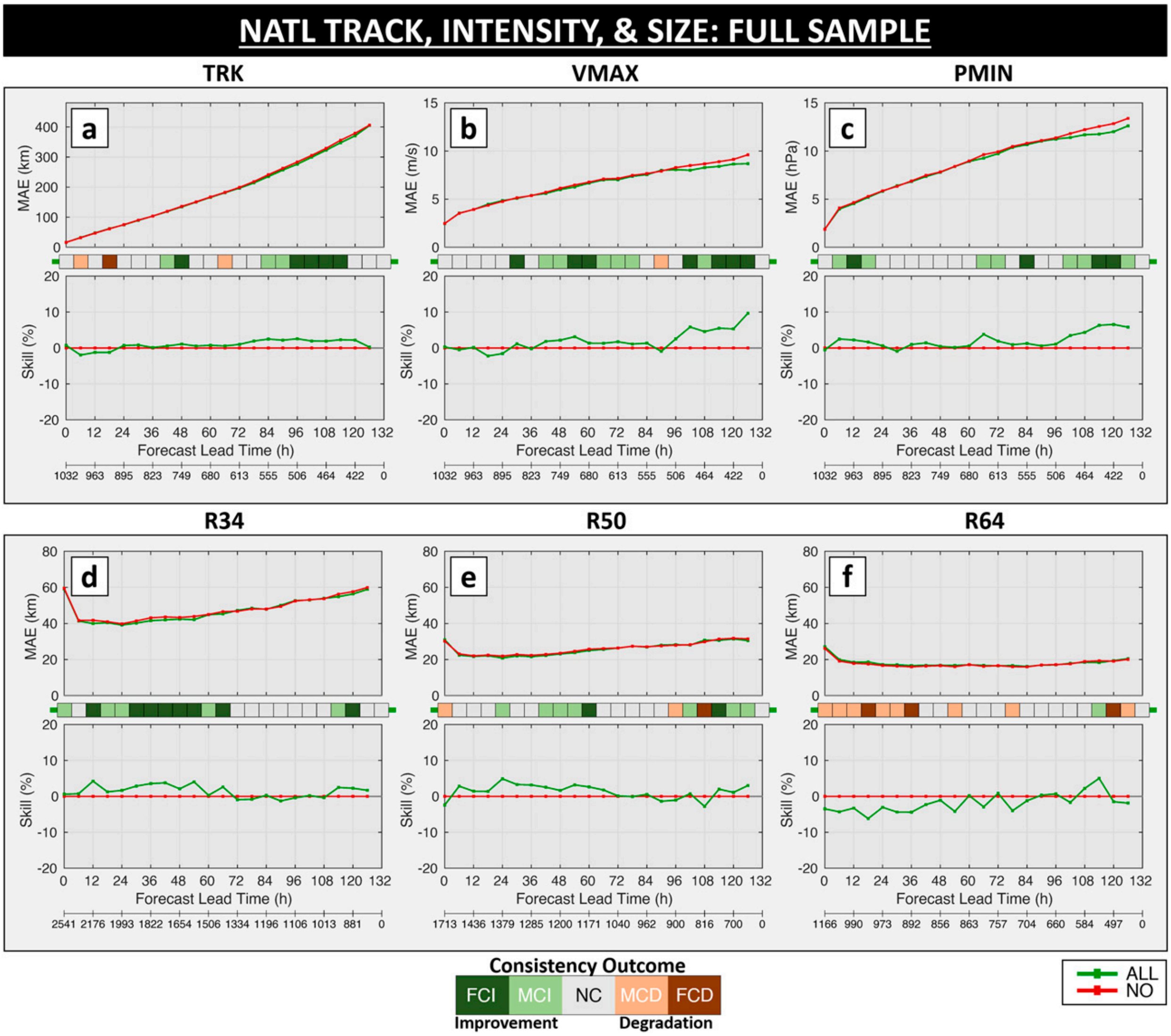


FIG. 6. The MAE and MAE skill for NATL TCs for the ALL (green) and NO (red) experiments for (a) track (TRK), (b) VMAX, (c) PMIN, (d) R34, (e) R50, and (f) R64. Shaded boxes between the MAE and MAE skill panels indicate forecast lead times where results were fully consistent, marginally consistent, or not consistent, based on the criteria shown in Fig. 5. The sample size is given below the *x* axis in each panel.

stratified by initial classification into four groups according to their Saffir–Simpson scale (Simpson and Saffir 1974) best track classification at 0 h: 1) tropical depression (TD; <17.5 m s^{-1}), 2) tropical storm (TS; ≥17.5 and <32.9 m s^{-1}), 3) category-1–2 hurricane (H12; ≥32.9 and <49.4 m s^{-1}), and 4) category-3–5 hurricane (H345; ≥49.4 m s^{-1}). Since TDs rarely had assimilated dropsonde observations, there are only nine verifiable forecasts at 0 h for TDs, resulting in too small of a sample size to conduct a robust analysis for this study. Finally, to explore the impact of dropsondes in an evolving vortex, OBS is also stratified by ongoing intensity change into three groups according to their ±6-h best track intensity change at 0 h: 1) intensifying (IN; >2.6 m s^{-1}), 2) steady state (SS; ≥−2.6 and ≤2.6 m s^{-1}), and 3) weakening (WK; <−2.6 m s^{-1}).

2) METRICS

For the full sample and each stratification, nine metrics are calculated for each variable, including the: 1) mean absolute error (MAE), 2) median absolute error (MDAE), 3) MAE skill, 4) MDAE skill, 5) frequency of superior performance (FSP; Velden and Goldenberg 1987; Goldenberg et al. 2015), 6) MAE TC-by-TC skill, 7) MDAE TC-by-TC skill, 8) TC-by-TC FSP, and 9) TC-by-TC sample size. Note that skill metrics use NO as a baseline. Descriptions of each metric can be found in Table 2.

Since nonnormal distributions can lead to misleading interpretations of MAE, this study assesses consistency between the MAE skill, MDAE skill, and FSP by using the "consistency

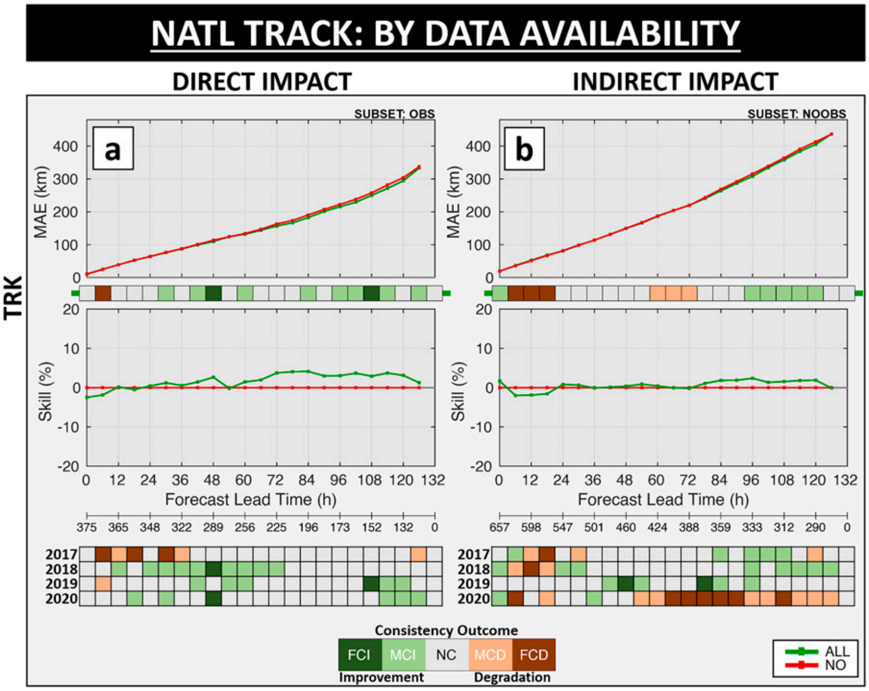


FIG. 7. As in Fig. 6, but for the (a) direct (OBS) and (b) indirect (NOOBS) impact of dropsonde observations on TC track forecasts. Consistency scorecards below both panels stratify (a) and (b) by year, respectively.

metric" introduced in Ditchek et al. (2023). This new metric objectively identifies forecast lead times that had fully consistent or marginally consistent improvement or degradation as defined in Fig. 5 by applying thresholds to the MAE skill, FSP, and MDAE skill. For more details on the consistency metric, see Ditchek et al. (2023).

For this study, discussion will generally focus on results that are fully consistent or marginally consistent across at least two

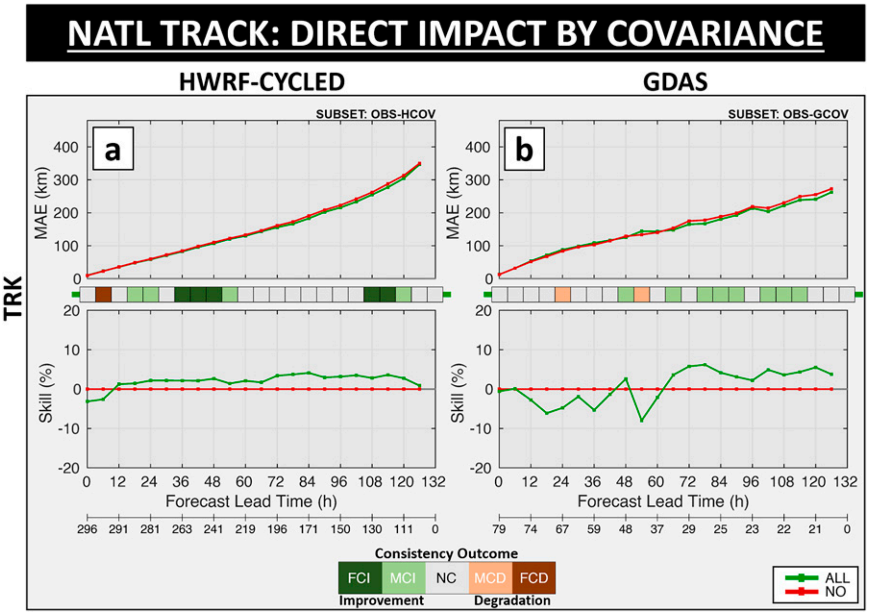


FIG. 8. As in Fig. 6, but for the impact of assimilated dropsonde observations on TC track forecasts for TCs that (a) used HWRF-cycled covariance (OBS-HCOV) and (b) used GDAS covariance (OBS-GCOV).

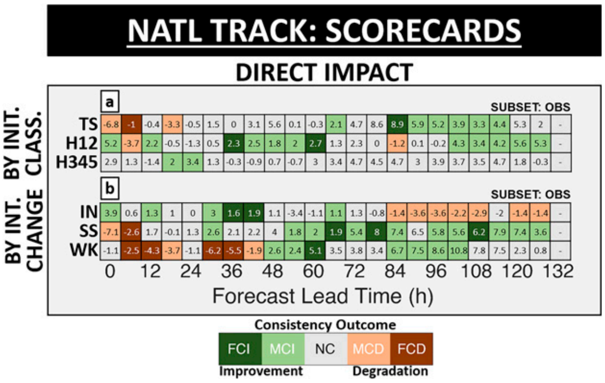


FIG. 9. Consistency scorecards for the direct impact of dropsondes on TC track forecasts, stratified by (a) initial classification (int. class.) and (b) ongoing intensity change (int. change). Note that tropical depression (TD) is not included in the consistency metric stratification in (a) due to the small sample size. MAE skill values are also included on each outcome, for reference. The sample sizes for TS, H12, H345, IN, SS, and WK at 0 h are 139, 121, 106, 128, 158, and 87, respectively, and at 120 h they are 34, 31, 62, 61, 54, and 16, respectively.

consecutive lead times. Note that even small improvements in MAE are notable if they occur consistently within a sample. Additionally, only MAE and MAE skill graphics with the augmented consistency metric and scorecard graphics displaying only consistency metric results will be shown. Still, all nine metrics were still used to analyze results to 1) quantify the magnitude of any observed improvement or degradation, 2) compare the errors between experiments, and 3) further identify drivers of distribution skewness.

3. NATL forecast performance

Statistics covering the entire sample indicate that dropsondes generally improved TC forecasts. For track (Fig. 6a), dropsondes improved forecasts at long lead times. While VMAX (Fig. 6b) in ALL improved from 24 h onward, PMIN (Fig. 6c) improved both on day 1 and at long lead times. Dropsondes also improved both R34 (Fig. 6d) and R50 (Fig. 6e) forecasts at short lead times. On the other hand, they degraded R64 forecasts (Fig. 6f). This degradation, hypothesized to result from observing system inadequacies in 2017, will be discussed in section 3c(2).

The rest of this section discusses results for track, intensity, and significant wind radii. For each variable, stratifying the full sample by data availability allows exploration of the direct (OBS) and indirect (NOOBS) impacts of dropsondes on TC forecasts. Both OBS and NOOBS are further stratified by year (e.g., OBS-2017). OBS is then stratified 1) by the choice of inner-core error covariance (e.g., OBS-HCOV and OBS-GCOV), 2) by the initial TC classification (e.g., OBS-TS), and 3) by the ongoing intensity change (e.g., OBS-SS). For more details on these stratifications, see section 2d(1).

a. Track

Dropsonde observations directly improved track forecasts, though their indirect impacts were less clear. Track in ALL directly improved by 1.7% on average and up to 4.1% (Fig. 7a), with fully consistent or marginally consistent results at about half of the lead times ≥ 30 h. This improvement was greater than that in the full sample (cf. Figs. 6a and 7a), indicating that dropsonde observations led to better track forecasts when directly assimilated into a TC's nested domains. Stratifying the results by year (Fig. 7a, bottom) reveals some interannual variability, which emphasizes the need for a large, multiyear sample. For example, ALL had degradation in 2017 at short lead times but improvement in 2018–20 at varying short and long lead times. The indirect impact on track forecasts (Fig. 7b) was weaker, had less consistency, and had even more interannual variability (Fig. 7b, bottom). Of particular note, dropsondes indirectly degraded 2020 TC track forecasts at most lead times after 48 h, a result not seen in other years.

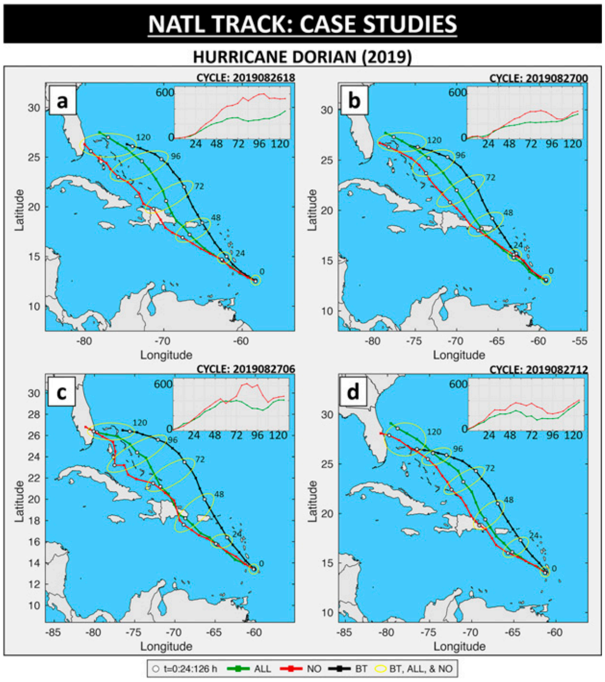


FIG. 10. A comparison of best track (BT; black), ALL (green), and NO (red) track forecasts from four individual forecasts in Hurricane Dorian (2019), where white dots are included every 24 h, for reference. For aesthetic purposes, every 24-h ellipses (yellow) encircle the BT, ALL, and NO forecast locations valid at the same time. Forecast lead times are indicated next to the ellipses for comparison to the associated insets, which depict the track MAE (km) of ALL and NO compared to the best track.

Using HWRF-cycled covariance appeared to influence the impact of dropsondes on TC track forecasts. Of note, OBS-HCOV entirely drove the short-lead-time improvement in Fig. 7a (cf. Figs. 7a and 8a). Such a strong disparity between OBS-HCOV and OBS-GCOV was not seen at longer lead times, where dropsondes positively impacted track forecasts in both cases. These results are qualitatively similar to L17, who found that using GDAS covariance likewise degraded track forecasts compared with using mesoscale covariance supplied by a cycled ensemble within HWRF.

The impact of dropsondes on track also depended considerably on the initial classification of the TC. For

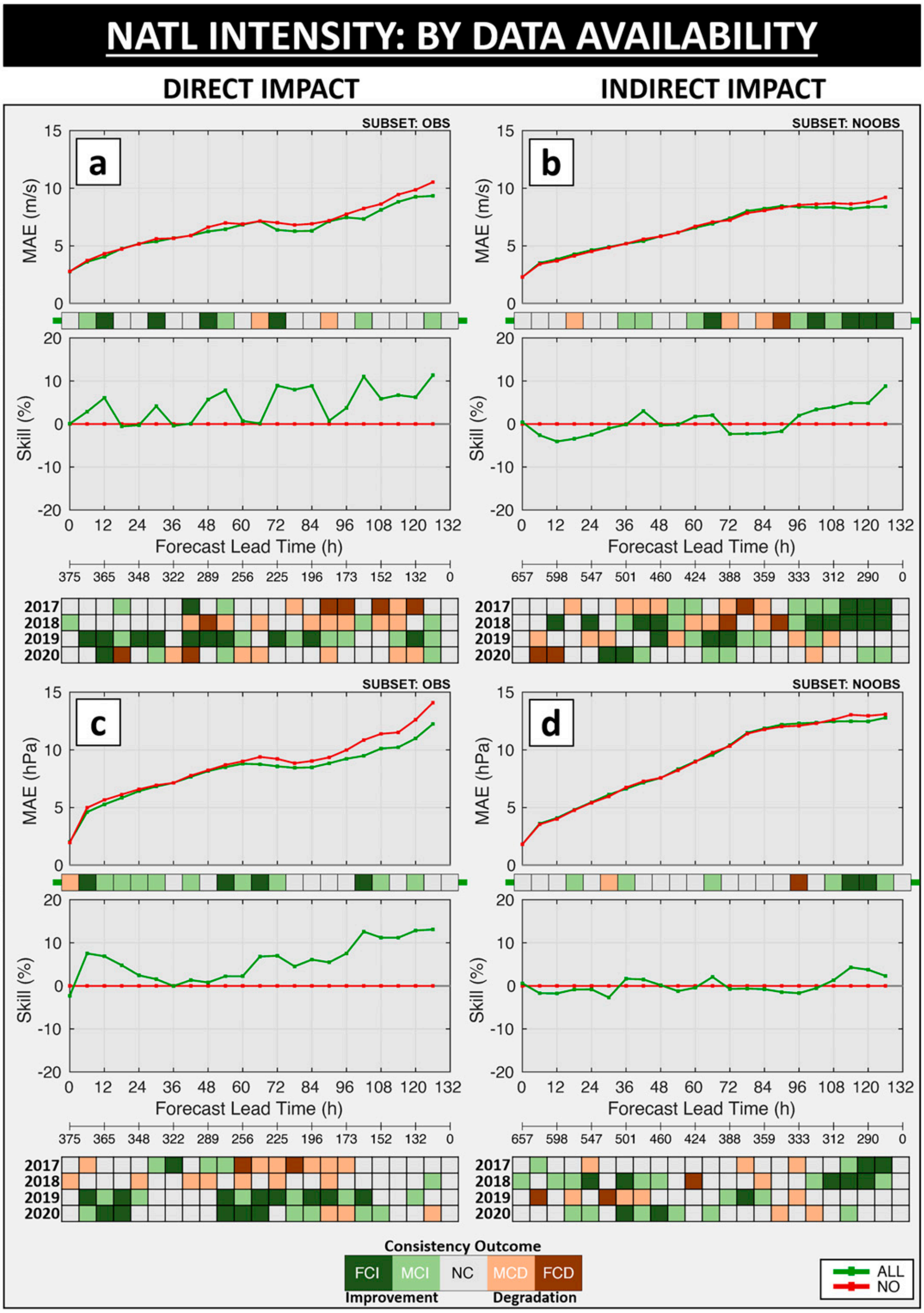


FIG. 11. As in Fig. 7, but for (a),(b) VMAX and (c),(d) PMIN forecasts.

example, sampling H12 with dropsondes led to more consistent benefits than sampling other intensity classifications (Fig. 9a). In particular, OBS-H12 had windows of at least marginally consistent improvement from 36 to 60 h and from 102 to 126 h. Meanwhile, dropsondes improved forecasts of OBS-TS only at long lead times and minimally impacted H345 forecasts. Dropsondes also improved track at most lead times ≥ 54 h in steady-state TCs (OBS-SS;

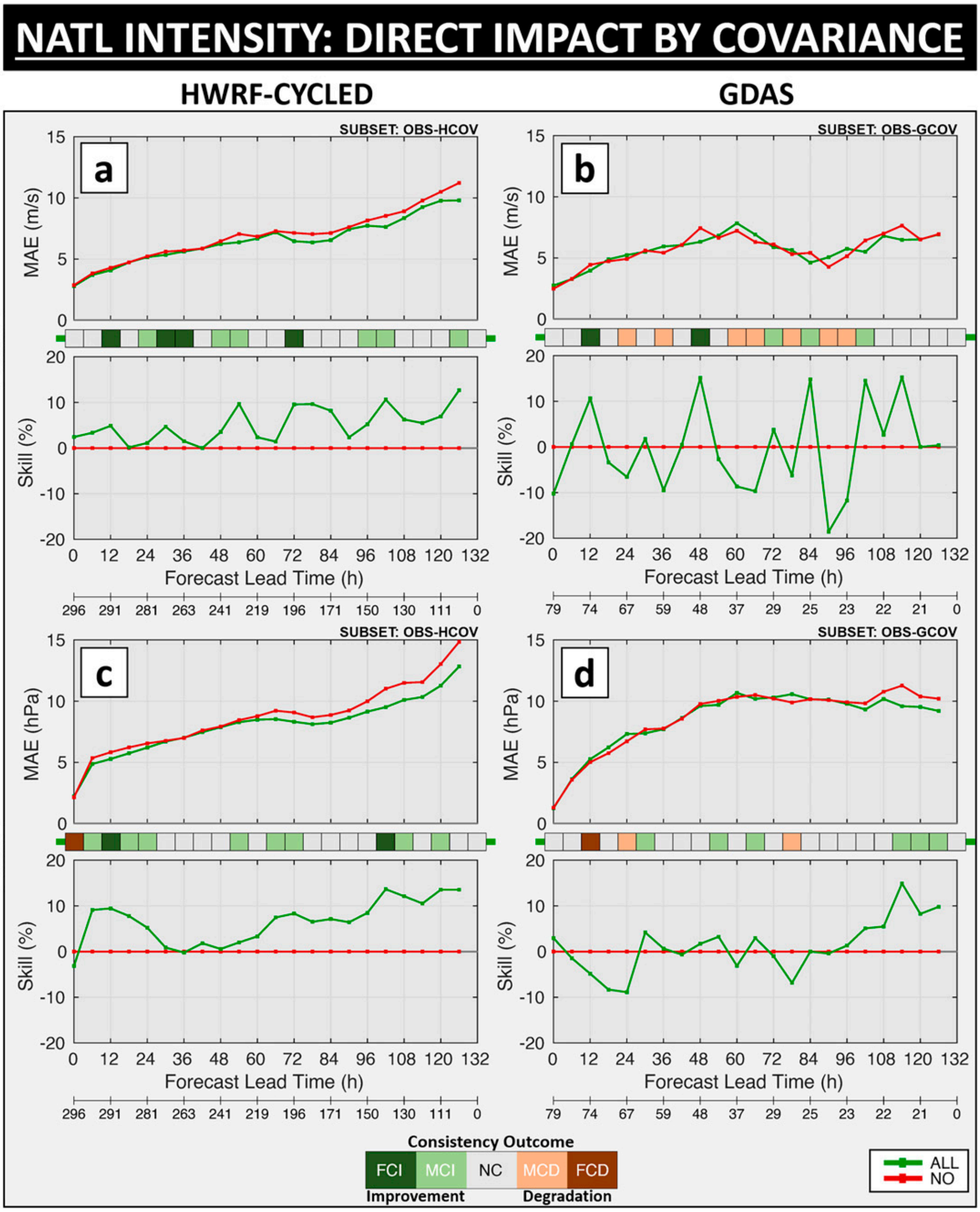
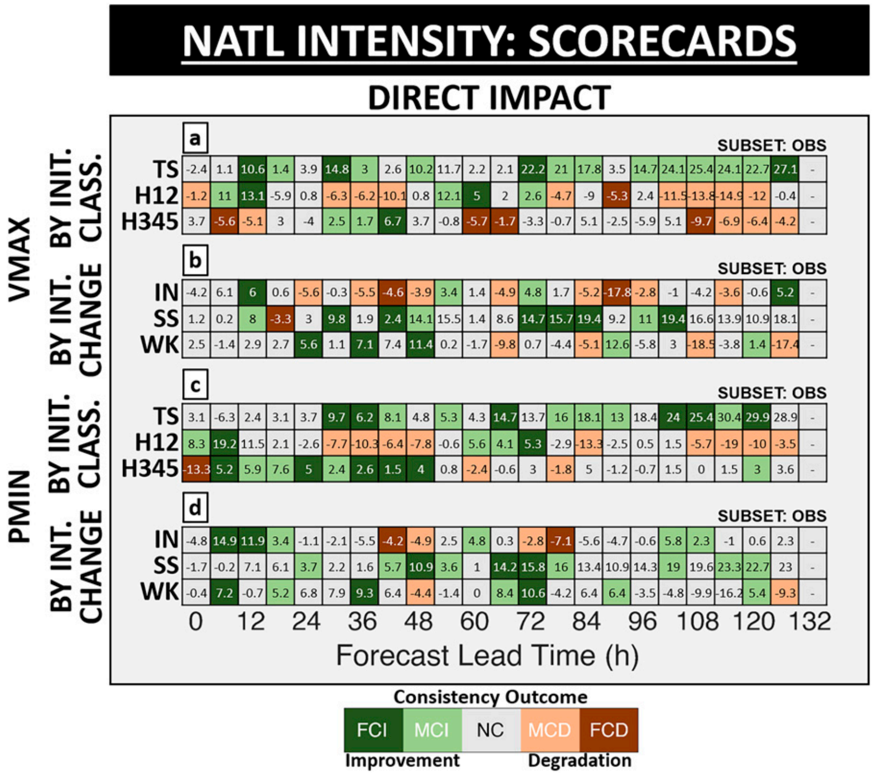


FIG. 12. As in Fig. 7, but for (a),(b) VMAX and (c),(d) PMIN for TCs that (left) used HWRF-cycled covariance (OBS-HCOV) and (right) used GDAS covariance (OBS-GCOV).

Fig. 9b) with some consistency, which is a generally more positive track impact than those seen for other intensity-change classifications.

The results in Fig. 9 reflect improvement in HWRF physics and DA since T18, though deficiencies still remain. For example, the degradations seen in T18 for stronger TCs did not occur here, and dropsondes consistently improved H12 track forecasts. Nevertheless, the degraded impact during intensity change suggests continued deficiencies in the DA system, as will be discussed later.

To demonstrate track improvements in a particular case where dropsondes played a key role in improving forecasts, Fig. 10 compares ALL, NO, and the observed best track during four individual forecasts of Hurricane Dorian (2019) during its TS phase. Note that dropsonde observations were assimilated to initialize each of the four forecasts, and all made use of the HWRF-cycled covariance. Further, Dorian's initial VMAX during these four forecasts remained constant at about 23 m s^{-1}. Thus, the OBS-HCOV, TS, and SS samples all include these forecasts, and the relevant



a																							
TS	-2.4	1.1	10.6	1.4	3.9	14.8	3	2.6	10.2	11.7	2.2	2.1	22.2	21	17.8	3.5	14.7	24.1	25.4	24.1	22.7	27.1	-
H12	-1.2	11	13.1	-5.9	0.8	-6.3	-6.2	-10.1	0.8	12.1	5	2	2.6	-4.7	-9	-5.3	2.4	-11.5	-13.8	-14.9	-12	-0.4	-
H345	3.7	-5.6	-5.1	3	-4	2.5	1.7	6.7	3.7	-0.8	-5.7	-1.7	-3.3	-0.7	5.1	-2.5	-5.9	5.1	-9.7	-6.9	-6.4	-4.2	-
b																							
IN	-4.2	6.1	6	0.6	-5.6	-0.3	-5.5	-4.6	-3.9	3.4	1.4	-4.9	4.8	1.7	-5.2	-17.8	-2.8	-1	-4.2	-3.6	-0.6	5.2	-
SS	1.2	0.2	8	-3.3	3	9.8	1.9	2.4	14.1	15.5	1.4	8.6	14.7	15.7	19.4	9.2	11	19.4	16.6	13.9	10.9	18.1	-
WK	2.5	-1.4	2.9	2.7	5.6	1.1	7.1	7.4	11.4	0.2	-1.7	-9.8	0.7	-4.4	-5.1	12.6	-5.8	3	-18.5	-3.8	1.4	-17.4	-
c																							
TS	3.1	-6.3	2.4	3.1	3.7	9.7	6.2	8.1	4.8	5.3	4.3	14.7	13.7	16	18.1	13	18.4	24	25.4	30.4	29.9	28.9	-
H12	8.3	19.2	11.5	2.1	-2.6	-7.7	-10.3	-6.4	-7.8	-0.6	5.6	4.1	5.3	-2.9	-13.3	-2.5	0.5	1.5	-5.7	-19	-10	-3.5	-
H345	-13.3	5.2	5.9	7.6	5	2.4	2.6	1.5	4	0.8	-2.4	-0.6	3	-1.8	5	-1.2	-0.7	1.5	0	1.5	3	3.6	-
d																							
IN	-4.8	14.9	11.9	3.4	-1.1	-2.1	-5.5	-4.2	-4.9	2.5	4.8	0.3	-2.8	-7.1	-5.6	-4.7	-0.6	5.8	2.3	-1	0.6	2.3	-
SS	-1.7	-0.2	7.1	6.1	3.7	2.2	1.6	5.7	10.9	3.6	1	14.2	15.8	16	13.4	10.9	14.3	19	19.6	23.3	22.7	23	-
WK	-0.4	7.2	-0.7	5.2	6.8	7.9	9.3	6.4	-4.4	-1.4	0	8.4	10.6	-4.2	6.4	6.4	-3.5	-4.8	-9.9	-16.2	5.4	-9.3	-

FIG. 13. As in Fig. 9, but for (a),(b) VMAX and (c),(d) PMIN.

results in Figs. 8a and 9 all suggested forecast improvement with dropsondes. Indeed, ALL outperformed NO at most lead times after 24 h in each forecast in Fig. 10. Most importantly, the four forecasts in NO suggested a landfall in southeastern Florida, while the forecasts in ALL correctly kept Dorian farther from the coast. These results highlight the importance of dropsonde sampling during the early part of Dorian's lifetime.

b. *Intensity*

Of all results in this study, the direct impact of dropsondes on VMAX forecasts perhaps best illustrates the need for a large, multiyear sample. While Fig. 11a shows that dropsondes directly improved VMAX on average, the improvement was quite inconsistent. A reduction of very large errors, mainly in individual forecasts from 2019 TCs with dropsondes, strongly influenced the MAE and MAE skill (Fig. 11a, bottom). In fact, OBS-2019 forecasts in ALL improved VMAX by 14.1% on average and up to 23.2% (not shown). On the other hand, VMAX forecasts in ALL degraded in OBS-2017-18 and OBS-2020. Thus, the indirect impact of dropsondes (NOOBS) actually drove consistency found in the full sample after 24 h (cf. Figs. 6b and 11a,b). Most notably, dropsondes indirectly improved VMAX by 4.7% on average for ≥96 h. This improvement was found during 2017, 2018, and partially during 2020 (Fig. 11b, bottom).

PMIN forecast improvement due to dropsondes varied less, and at least marginally consistent improvement was found in ALL at most lead times (Fig. 11c). Between 6 and 18 h, PMIN forecasts in ALL improved by 6.4% on average and demonstrated some consistency in 2019–20 (Fig. 11c, bottom). Improvement was greater than that in the full sample (cf. Figs. 6c and 11c), indicating that dropsondes lead to better 6–18-h PMIN forecasts in individual forecasts where they were assimilated. As with VMAX, forecasts from TCs in 2019 with dropsondes drove the magnitude of PMIN improvements at longer lead times (Fig. 11c, bottom). During that year, PMIN in ALL improved by 16.4% on average and up to 27.1% (not shown). On the other hand, at least marginally consistent degradation was found for OBS-2017-18 in ALL. The indirect impact of dropsondes on PMIN forecasts was mostly neutral, but it did drive the consistency found in the full sample at longest lead times (cf. Figs. 6c and 11c,d). In particular, TCs from 2017 and 2018 helped improve PMIN in ALL by 2.9% on average at lead times ≥ 108 h.

As in L17, choice of inner-core covariance also strongly influenced the impact of dropsondes on intensity. VMAX improvements for ALL in Fig. 11a only occurred when using HWRF-cycled covariance (OBS-HCOV; Fig. 12a). In fact, dropsondes degraded VMAX in OBS-GCOV with marginal consistency across multiple lead times (Fig. 12b). For PMIN, the impact of dropsondes was again more positive in OBS-HCOV, particularly during the first 24 h. The consistency seen in the OBS-HCOV stratification (Fig. 12c) was not present in OBS-GCOV (Fig. 12d). Finally, contrary to VMAX, at the longest lead times dropsondes improved PMIN forecasts with at least marginal consistency in both OBS-GCOV and OBS-HCOV.

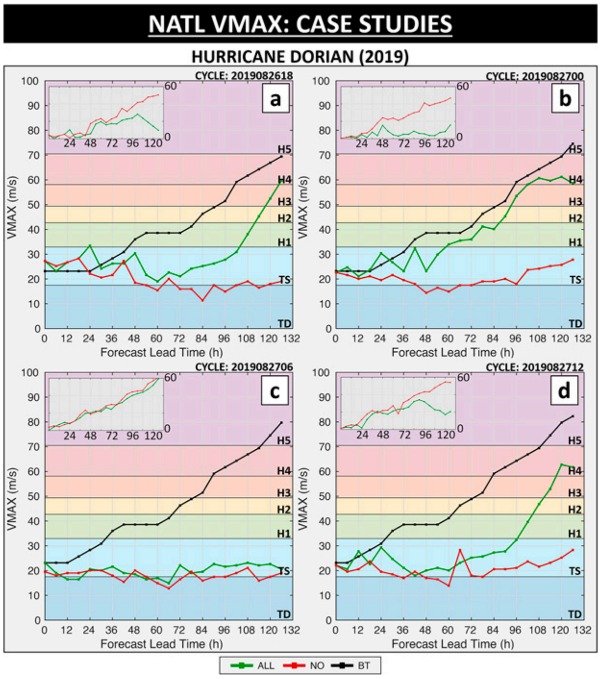


FIG. 14. As in Fig. 10, but for VMAX, where shading from dark blue to magenta corresponds to the TC classification from TD through category-5 hurricane (H5).

Similar to the results for track, the impacts of dropsondes on the intensity forecast depended on the initial classification. In particular, sampling TS with dropsondes led to more intensity forecast improvement than sampling other classifications (Fig. 13). For OBS-TS, dropsondes improved both VMAX and PMIN between 10% and 30% with at least marginal consistency, though improvement ≥ 36 h was mainly due to individual forecasts from 2019 TCs (not shown). Dropsondes also improved PMIN forecasts in OBS-H345 between 6 and 48 h, which could be valuable for landfalling cases since PMIN is better correlated with damage than VMAX (e.g., Chavas et al. 2017; Klotzbach et al. 2020).

As with track, most of the intensity improvement in OBS occurred when dropsondes sampled steady-state TCs (OBS-SS; Figs. 13b,d). For ALL, both VMAX and PMIN forecasts generally improved between 10% and 20% in OBS-SS ≥ 36 h with some consistency. Further, short term improvements were found for PMIN in intensifying TCs (OBS-IN).

As alluded to previously, the degraded impact of dropsondes when intensity changes suggests a deficiency in the DA system. In particular, Lu et al. (2017b) and Davis et al. (2021) both showed that even an advanced configuration with HWRF-cycled covariance can have problems when GSI uses 3DEnVAR. Issues arise because HWRF uses a 6-h window for DA, and 3DEnVAR does not account for the time evolution of error covariance. For steady-state systems, the assumption of temporally invariant covariance within the DA window might suffice, but for situations when the inner core evolves it does not. Indeed, analyses and forecasts initialized with inner-core DA improve when GSI uses 4DEnVAR or when 3DEnVAR is cycled with hourly updates (Lu et al. 2017b; Davis et al. 2021).

To illustrate how the above results relate to an example, Fig. 14 shows the evolution of VMAX forecasts and errors for the same four individual forecasts given in Fig. 10 for Hurricane Dorian (2019). Note that the OBS-HCOV (Fig. 12a), OBS-TS (Fig. 13a), and OBS-SS (Fig. 13b) stratifications all include these forecasts, and each suggests forecast improvement. Indeed, ALL outperformed NO at most lead times of every forecast in Fig. 14, and the improvements were substantial in all but Fig. 14c. Most strikingly, NO consistently kept Dorian as a TS, while ALL captured Dorian's intensification to a category-4 hurricane in three of the four representative forecasts. In the forecast that did not forecast Dorian's intensification (Fig. 14c), the corresponding track forecast (Fig. 10c) had Dorian making landfall in Hispaniola. This is likely the reason for the lack of intensification found. Yet, ALL still outperformed NO at most lead times, particularly at long lead times. These results again highlight the importance of dropsonde sampling for Dorian's forecasts early in its lifetime.

c. *Significant wind radii*

This section examines the impacts of dropsondes on TC significant-wind-radii forecasts. Impacts on the outer wind radii (R34 and R50) are discussed separately from R64 due to distinctly different results for those metrics. Probable reasons for those differences are discussed below.

1) R34 AND R50

Among all variables evaluated, R34 and R50 forecasts saw the most consistent impact from dropsonde sampling. Dropsondes directly improved R34 forecasts 2.4% on average and up to 7.5% (Fig. 15a) as well as R50 forecasts 3.3% on average and up to 7.2% (Fig. 15c). These improvements were at least marginally consistent at most lead times and were also greater than those found in the full sample (Figs. 6d,e). Additionally, ALL improved upon NO every year in the sample (i.e., interannual stationarity; Figs. 15a,c, bottom). Dropsondes also indirectly improved R34 at most lead times through 66 h in ALL with at least marginal consistency (Fig. 15b). While most improvement came from 2017 (Fig. 15b, bottom), various lead times during 2018–20 also saw fully consistent improvements (Fig. 15b, bottom). Finally, the indirect impact of dropsondes on R50 provides yet another example of the importance of a multiyear sample. In particular, R50 in ALL consistently degraded at most lead times > 60 h in 2020, though such a strong signal did not occur in other years.

Further stratifying the results emphasizes the hypothesis that the details of DA in HWRF strongly influence the impact of dropsondes. As with track and intensity, most of the direct benefits of dropsondes to forecasts of R34 and R50 occurred in OBS-HCOV (Fig. 16). The disparity of impact between OBS-HCOV and OBS-GCOV highlights the importance of using appropriate mesoscale covariance for forecasts of TC outer wind radii. Note that these results cannot be compared with L17 since they did not assess the differences in TC outer wind radii forecasts when using GDAS compared to HWRF-cycled covariance.

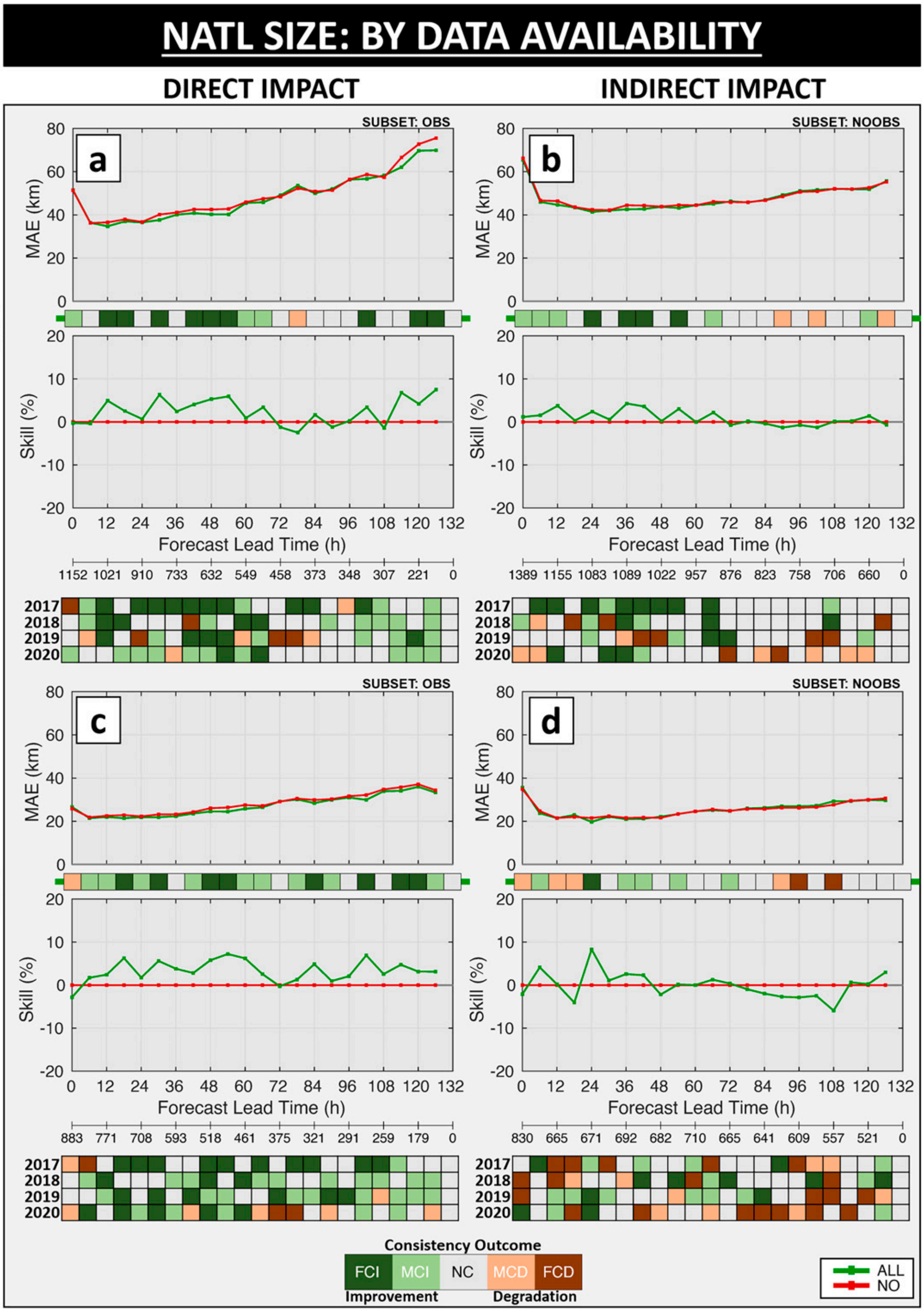


FIG. 15. As in Fig. 7, but for (a),(b) R34 and (c),(d) R50 forecasts.

Stratifying OBS by initial classification reveals that sampling hurricanes improved both R34 and R50 forecasts the most, with H12 having consistent results over a larger window than H345 (Fig. 17). For both variables, multiple lead times in ALL saw fully consistent improvement of more than 5%. R34 and R50 forecasts of OBS-TS also benefited from dropsonde sampling, though mainly at longer lead times compared to hurricanes (i.e., ≥90 h). While these results cannot be compared with T18 since they did

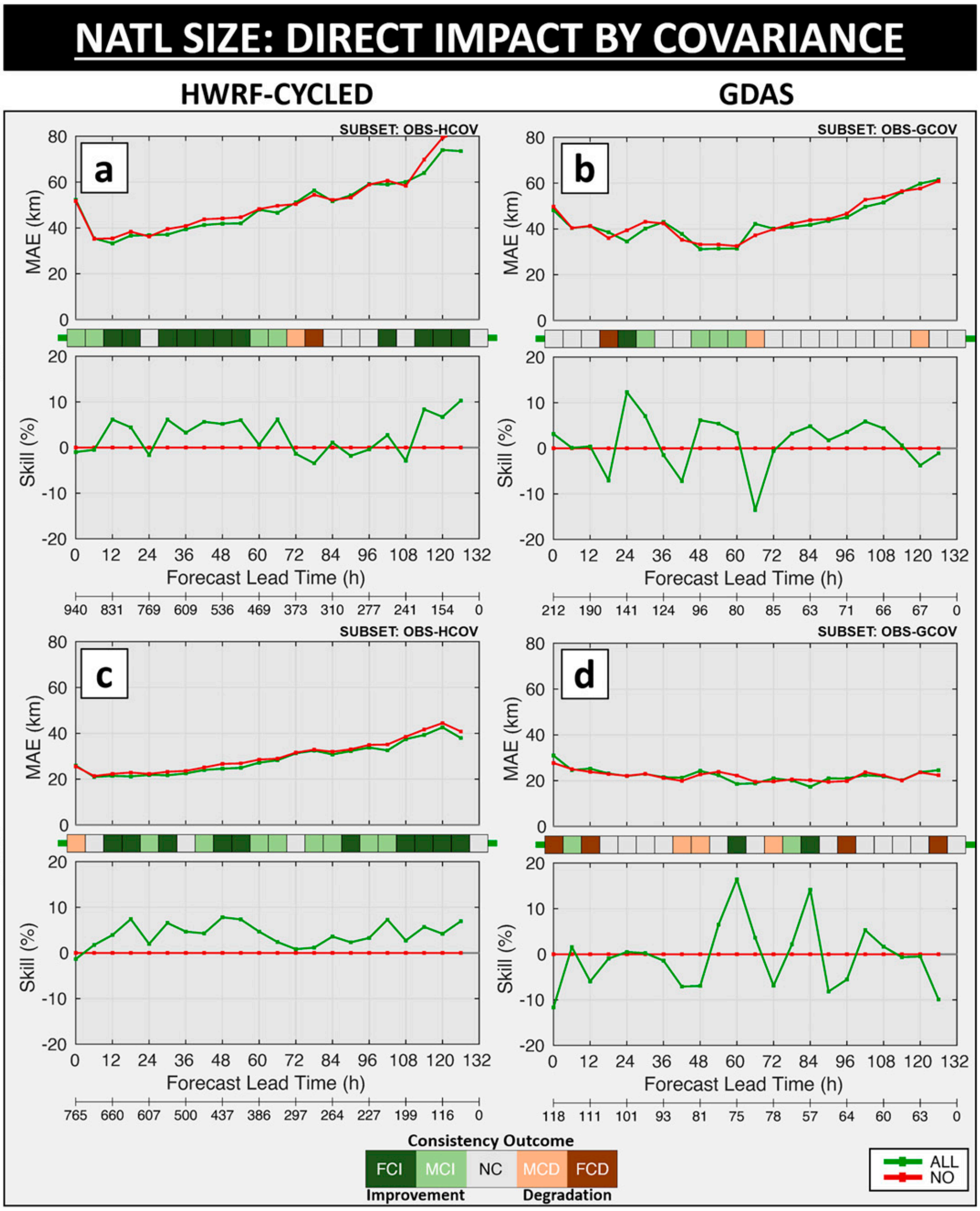


FIG. 16. As in Fig. 7, but for (a),(b) R34 and (c),(d) R50 for TCs that (left) used HWRF-cycled covariance (OBS-HCOV) and (right) used GDAS covariance (OBS-GCOV).

not examine the variability of TC outer-wind-radii forecasts as a function of TC intensity, the strong positive benefits to TC outer wind radii seen here for hurricanes are encouraging. Similar to the results for track and intensity forecasts, dropsondes were most valuable for TC outer wind radii forecasts when sampling steady-state TCs (Figs. 17b,d). The diminished benefits for TCs undergoing intensity change further suggests that the 6-hourly 3DEnVAR DA configuration hinders greater positive impacts.

A forecast of Hurricane Dorian (2019) initialized at 1800 UTC 29 August 2019 demonstrates how R50 improvements pertain to an individual case. This particular forecast was chosen since the track and VMAX for ALL and NO were similar (Figs. 18a,b), which eliminates a source of bias in the outer-wind-radii comparison. Note that this forecast is included in the Fig. 16c assessment as well as in the OBS-IN and OBS-H12 stratifications in Figs. 17c and 17d since Dorian was an intensifying category-1 hurricane and had assimilated dropsonde observations. From those results, improvement would be expected, with more fully consistent improvement found for forecasts that use HWRF-cycled covariance. Indeed, ALL outperformed NO through about

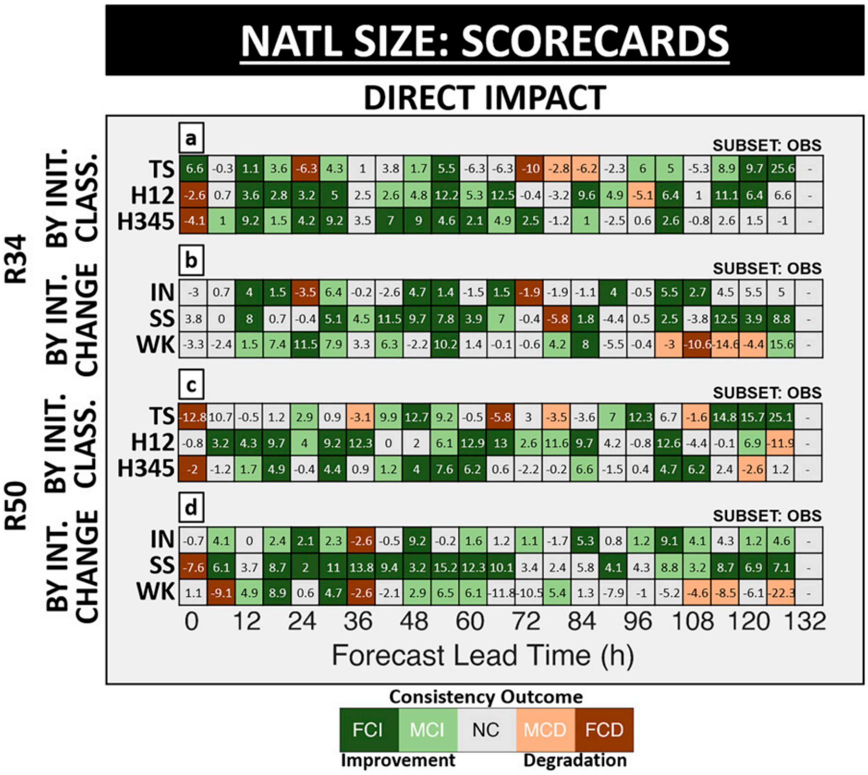


FIG. 17. As in Fig. 9, but for (a),(b) R34 and (c),(d) R50. The sample sizes for R34 for TS, H12, H345, IN, SS, and WK at 0 h are 334, 418, 400, 411, 482, and 259, respectively, and at 120 h are 45, 45, 126, 114, 86, and 19, respectively. For R50 for TS, H12, H345, IN, SS, and WK, the sample sizes at 0 h are 95, 388, 400, 325, 345, 213, respectively, and at 120 h are 33, 36, 109, 101, 62, 16, respectively.

36–48 h in Figs. 18c–f, which compares the R50 values in each quadrant of Dorian.

Figures 18g and 18h demonstrate how this result applies to the two-dimensional 10-m wind field using the 24-h forecast lead time as a representative example. For each experiment, the figure overlays 10-m winds with the observed R50 extent as well as the ALL and NO R50 extent, respectively. In all quadrants, R50 in ALL more closely matched the best track value. This type of TC outer wind radii improvement can be particularly valuable for landfalling cases.

2) R64

While dropsondes overall improved forecasts of both R34 and R50 at short lead times (Figs. 6d,e), Fig. 6f indicates that dropsondes degraded R64 forecasts. The source of that degradation was TCs from 2017 that directly assimilated dropsonde observations (OBS-2017; Fig. 19a, bottom and Fig. 19c). Fully consistent or marginally consistent degradation occurred at several lead times for OBS-2017, with degradation of 6.9% on average and up to 23.2%. Similar degradation was not found for 2018–20 (Fig. 19a, bottom) or for the indirect impact of dropsondes on TC forecasts of R64 (Fig. 19b). Additionally, the negative 2017 results appeared in both the OBS-HCOV and OBS-GCOV samples (not shown).

Preliminary analysis suggests that insufficient dropsonde sampling of the R64 region[9] could have caused the degradation in OBS-2017. Figures 19c–f depicts the number of assimilated dropsonde observations within the R64 region (gray shading) for each year, respectively. Notice that there were on average three times the number of assimilated dropsonde observations within the R64 region in 2018–20 (Figs. 19d–f) compared to 2017 (Fig. 19c). The most likely reason for this disparity was the several dropsonde-sampling changes implemented after 2017 (see section 2b and Fig. 3). Thus, the better direct sampling of the R64 region by dropsondes in 2018–20, particularly the inner and near-core region (≤150 km), seems to have prevented similar day-1 degradation as found in 2017. Possible reasons for such sensitivity to sampling strategy include model biases and inadequate mesoscale error covariance within the DA system (e.g., Lu et al. 2017b, L17), where a strong horizontal gradient in sampling could lead to larger errors. Confirming the above hypothesis would require additional data-denial experiments for 2018–20, which are outside the scope of this study.

[9] The R64 region is defined as the mean ± the standard deviation of the best track R64 estimates for a given year.

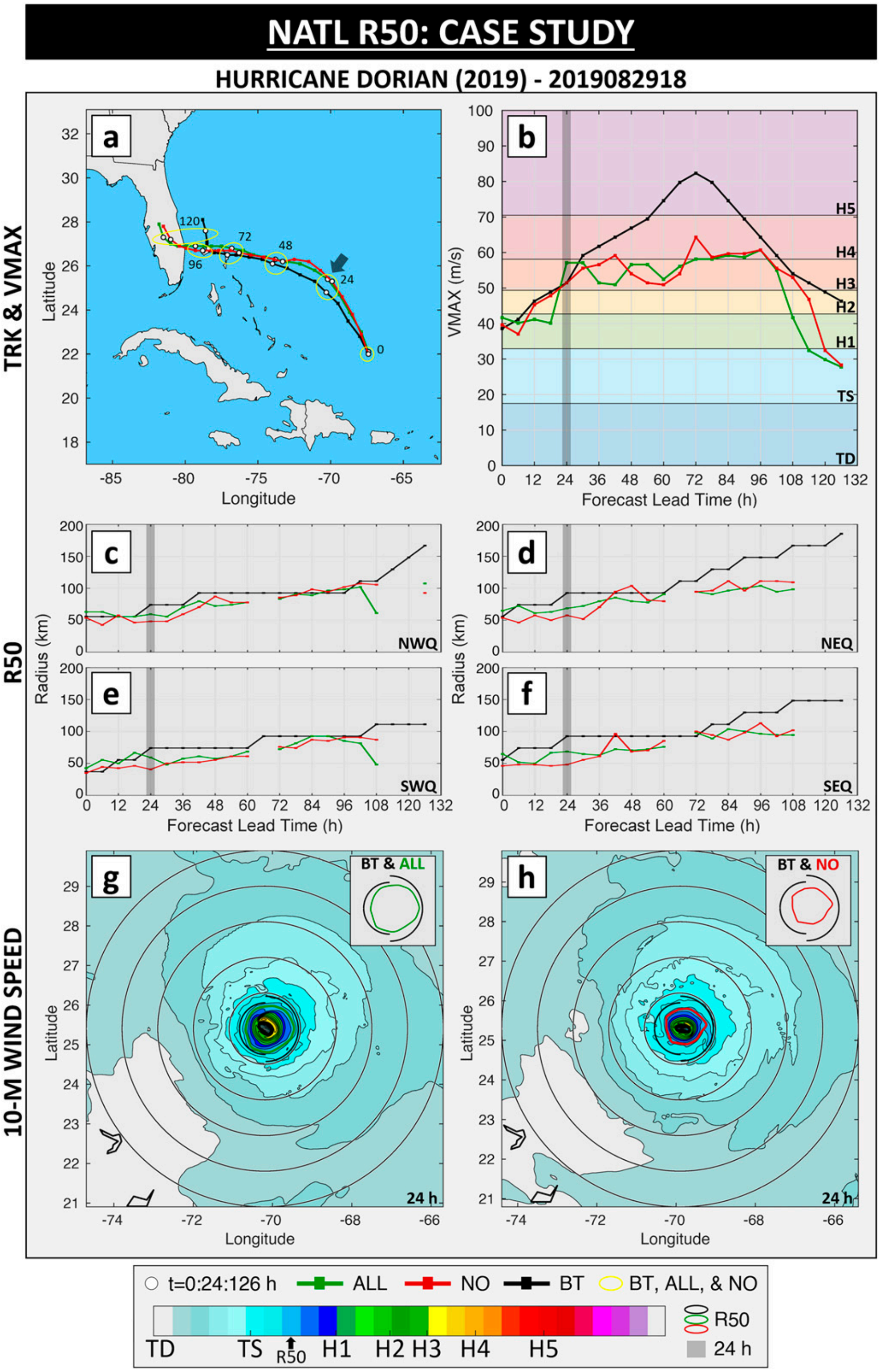


FIG. 18. For one individual forecast of Hurricane Dorian (2019) (a),(b) as in Figs. 10 and 14; (c)–(f) a comparison of best track (BT; black), ALL (green), and NO (red) R50 forecasts; and (g),(h) the 10-m wind speed at 24 h (shaded) in ALL and NO, respectively, with the BT, ALL, and NO R50 extents overlaid as well as in the inset. Note that gray shading in (a)–(f) indicates the 24-h forecast lead time.

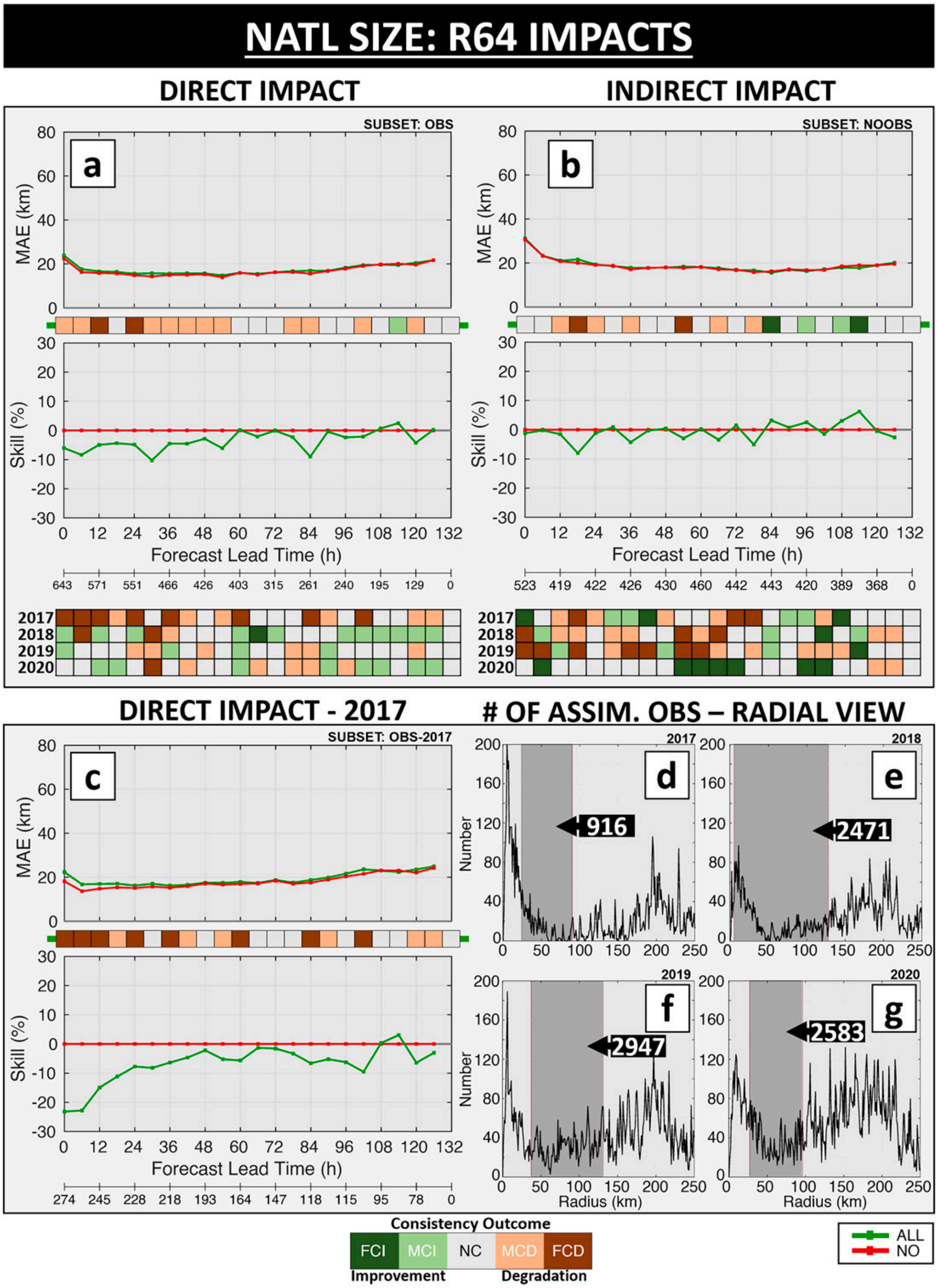
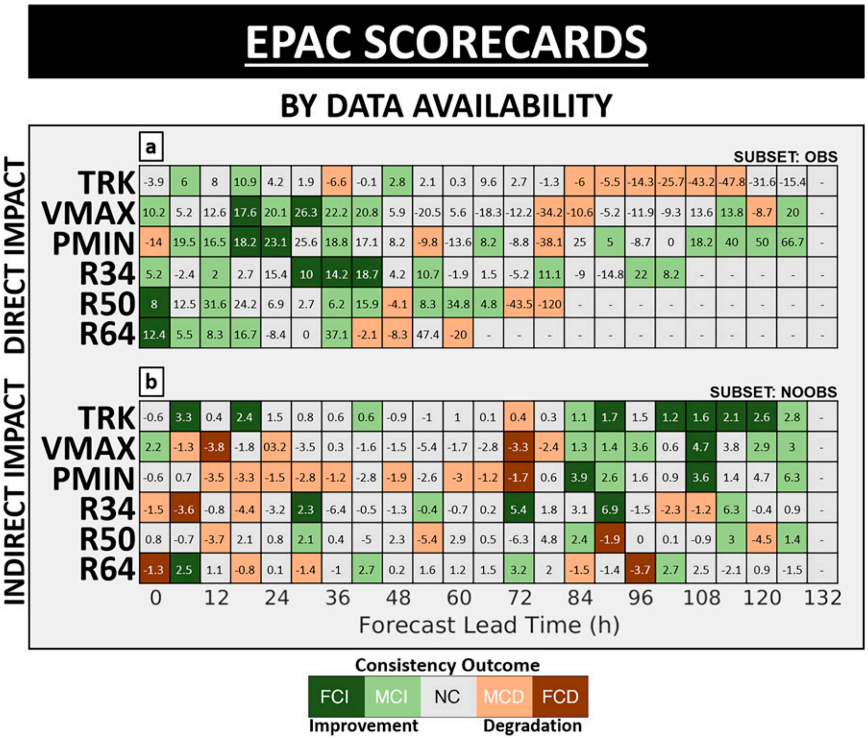


FIG. 19. The impact of dropsonde on R64 forecasts, including (a),(b) as in Fig. 7, but for R64, (c) the R64 MAE and MAE skill for OBS-2017, and (d)–(g) the number of assimilated dropsonde observations from 0 to 250 km with the R64 window (i.e., the mean ± the standard deviation of the best track R64 observations) in gray shading.

4. EPAC forecast performance

Though this study focuses on active NATL periods, use of HB20 allows for assessment of impacts for EPAC TCs as well (Table 1). As the EPAC sample has about half the size of NATL at 0 h and about a third at 120 h, no additional stratifications will be shown, and forecast lead times with sample sizes of <10 (i.e., in OBS > 72 h) will not be assessed. Note that stratifying OBS by D03 covariance reveals similar results to those found in the NATL, though the sample size is tiny.



a	0		12		24		36		48		60		72		84		96		108		120		132
TRK	-3.9	6	8	10.9	4.2	1.9	-6.6	-0.1	2.8	2.1	0.3	9.6	2.7	-1.3	-6	-5.5	-14.3	-25.7	-43.2	-47.8	-31.6	-15.4	-
VMAX	10.2	5.2	12.6	17.6	20.1	26.3	22.2	20.8	5.9	-20.5	5.6	-18.3	-12.2	-34.2	-10.6	-5.2	-11.9	-9.3	13.6	13.8	-8.7	20	-
PMIN	-14	19.5	16.5	18.2	23.1	25.6	18.8	17.1	8.2	-9.8	-13.6	8.2	-8.8	-38.1	25	5	-8.7	0	18.2	40	50	66.7	-
R34	5.2	-2.4	2	2.7	15.4	10	14.2	18.7	4.2	10.7	-1.9	1.5	-5.2	11.1	-9	-14.8	22	8.2	-	-	-	-	-
R50	8	12.5	31.6	24.2	6.9	2.7	6.2	15.9	-4.1	8.3	34.8	4.8	-43.5	-120	-	-	-	-	-	-	-	-	-
R64	12.4	5.5	8.3	16.7	-8.4	0	37.1	-2.1	-8.3	47.4	-20	-	-	-	-	-	-	-	-	-	-	-	-

b	0		12		24		36		48		60		72		84		96		108		120		132
TRK	-0.6	3.3	0.4	2.4	1.5	0.8	0.6	0.6	-0.9	-1	1	0.1	0.4	0.3	1.1	1.7	1.5	1.2	1.6	2.1	2.6	2.8	-
VMAX	2.2	-1.3	-3.8	-1.8	03.2	-3.5	0.3	-1.6	-1.5	-5.4	-1.7	-2.8	-3.3	-2.4	1.3	1.4	3.6	0.6	4.7	3.8	2.9	3	-
PMIN	-0.6	0.7	-3.5	-3.3	-1.5	-2.8	-1.2	-2.8	-1.9	-2.6	-3	-1.2	-1.7	0.6	3.9	2.6	1.6	0.9	3.6	1.4	4.7	6.3	-
R34	-1.5	-3.6	-0.8	-4.4	-3.2	2.3	-6.4	-0.5	-1.3	-0.4	-0.7	0.2	5.4	1.8	3.1	6.9	-1.5	-2.3	-1.2	6.3	-0.4	0.9	-
R50	0.8	-0.7	-3.7	2.1	0.8	2.1	0.4	-5	2.3	-5.4	2.9	0.5	-6.3	4.8	2.4	-1.9	0	0.1	-0.9	3	-4.5	1.4	-
R64	-1.3	2.5	1.1	-0.8	0.1	-1.4	-1	2.7	0.2	1.6	1.2	1.5	3.2	2	-1.5	-1.4	-3.7	2.7	2.5	-2.1	0.9	-1.5	-

FIG. 20. As in Fig. 9, but for the (a) direct and (b) indirect impact of dropsondes on EPAC TC track, intensity, and significant wind radii forecasts. The sample sizes for TRK, VMAX, PMIN, R34, R50, and R64 for OBS (NOOBS) at 0 h are 37, 37, 37, 130, 102, and 56 (498, 498, 498, 1374, 812, and 581), respectively, and at 120 h are 3, 3, 3, 0, 0, and 0 (144,144, 144, 425, 270, and 121), respectively.

Since relatively few dropsondes were deployed in EPAC TCs in the periods covered by this study, the sample size for EPAC OBS is quite small—only 8% of individual EPAC forecasts had assimilated dropsonde observations (Figs. 4c–f). Additionally, in those TCs where dropsonde observations were assimilated, EPAC TCs had fewer total assimilated dropsonde observations than NATL TCs (Fig. 4). Thus, NOOBS constitutes nearly the entire full sample. For brevity, the full sample will not be shown or discussed. Further, given the basin-scale, multistorm nature of HB20, any impacts seen in NOOBS are likely due to dropsonde assimilation in NATL TCs. While far apart, remote impacts to TCs across basins in HB20 has been previously explored and documented by Alaka et al. (2022).

The impact of dropsondes on EPAC TC forecasts varied in the direct (OBS; Fig. 20a) and indirect (NOOBS; Fig. 20b) samples. Dropsonde observations directly improved EPAC TC intensity and significant-wind-radii forecasts through 42 h with some consistency. Thereafter, results were mixed with a very small sample size. Significant-wind-radii improvements found were qualitatively similar to those found in NATL TCs (Fig. 16). Meanwhile, dropsonde observations indirectly improved EPAC track and intensity forecasts at long lead times with at least marginal consistency. This suggests that dropsondes assimilated in NATL TCs likely impacted EPAC forecasts. Note that these intensity results resemble the NOOBS results from the NATL sample, where dropsondes indirectly improved forecasts at long lead times. A notable exception is that dropsonde observations degraded PMIN in NOOBS with some consistency through 72 h.

5. Conclusions

This study thoroughly quantifies the impact of dropsondes released into TCs during active NATL periods within the 2017–20 hurricane seasons. To do so, it uses the 2020 version of the basin-scale, multistorm configuration of HWRF (HB20) to conduct two experiments: 1) one that assimilated dropsonde observations (ALL) and 2) one that did not (NO). These experiments included 634 cycles resulting in 2139 individual forecasts covering 92 TCs, 41 of which had assimilated dropsonde observations. Since HB20 has multiple interacting, high-resolution movable nests that track up to five TCs simultaneously within a large, static parent domain, assimilating dropsondes observation in any TC impacted forecasts of all TCs in the basin.

The performance of each experiment is evaluated by verifying forecasts of track, VMAX, PMIN, R34, R50, and R64 against the final NHC best track available from NHC, following standard NHC forecast verification procedures (Cangialosi 2022). Impacts found in the full sample are assessed by examining a variety of metrics and by taking various stratifications of the full sample to better understand the direct (OBS) and indirect (NOOBS) impact of dropsondes on TC forecasts in total and by year. OBS is further stratified by covariance used, by initial classification, and by ongoing intensity change [section 2d(1)]. To guide interpretation of results, this paper uses the consistency metric introduced and detailed in Ditchek et al. (2023). That metric objectively evaluates the evolution of forecast errors as a function of lead time based on thresholds applied to three metrics: MAE skill, MDAE skill, and the frequency of superior performance (FSP).

Before discussing results on the impact of dropsondes on TC forecasts, there are three takeaways learned that are applicable not only to studies that assess the impact of observing systems on TC forecasts, but also to those studies that assess the impact of new modeling systems or model upgrades on TC forecasts. First and foremost, this work highlights the tremendous importance of conducting large sample, multiyear studies, given year-to-year, TC-to-TC, and even forecast-to-forecast variability. The results here strongly indicate that running observation systems tests over time periods shorter than a few hurricane seasons yields insufficiently large samples to generate robust assessments. As studies typically do not have the resources needed for such a comprehensive assessment, sample-size caveats should be emphasized, and results from small-sample studies should not be generalized. Second, by using a large sample, stratifications with meaningful sample sizes can help to diagnose what drives any observed impacts in the full sample. Note that there are a number of stratifications that can be taken beyond those presented in this study. For this work, the stratifications reveal areas that can be improved to maximize data impact. Finally, even with large sample sizes, relying on commonly used metrics like the MAE and/or MAE skill could lead to misleading conclusions if the error distributions are skewed, which often occurs. Thus, using metrics beyond the MAE and MAE skill to understand the distribution of the errors or using a metric like the consistency metric (Ditchek et al. 2023) is needed to prevent reaching misleading conclusions.

By using a large sample, taking multiple stratifications, and analyzing results with the consistency metric, this study marks the most comprehensive assessment of the impact of dropsondes on NWP forecasts of TC track, intensity, and significant wind radii to date. In doing so, it is also the most comprehensive assessment of any airborne observing system on TC forecasts to date. The main focus of this paper (section 3) was on NATL[10] forecast performance. To aid in summarizing the main takeaways from this work, Fig. 21 shows the scorecards for all forecast variables assessed.

While dropsondes both directly and indirectly impacted NATL TC forecasts, direct sampling of TCs with dropsondes clearly yielded the greatest forecast improvements (cf. Figs. 21a,b). Particularly notable was the impact of dropsondes on TC outer-wind-radii forecasts, since improving those forecasts can lead to more effective TC-hazard forecasts, including those for storm surge, wind, rainfall, and associated freshwater flooding. These more accurate forecasts can lead to more refined and effective watches and warnings and also enable emergency managers and local officials to prepare and execute better preparation, mitigation, and evacuation strategies during an impending TC event. This study also found degradation in R64 forecasts during the 2017 hurricane season. This degradation probably occurred due to a lower number of inner- and near-core region (≤150 km) observations in 2017 than in subsequent years. This strongly suggests that sampling the TC inner and near core with dropsondes is critical for improving forecasts of the inner-core size.

The direct impact of dropsondes on TC forecasts detailed above is heavily dependent on DA quality, as nearly all of the benefits occurred when HWRF-cycled covariance was used (cf. Figs. 21c,d). In fact, since using GDAS covariance led to generally neutral to negative impacts, improvements were even more pronounced in Fig. 21c than in Fig. 21a. While an additional experiment which disables the use of HWRF-cycled covariance is needed to concretely quantify its impact (i.e., as done in L17), results here do suggest that using mesoscale error covariance native to HB20 is a vital part of the DA system. This is broadly consistent with the findings of L17, who demonstrated that using GDAS covariance to assimilate inner-core data results in strongly asymmetric and physically unrealistic analysis increments that can degrade all aspects of a forecast. T18 further demonstrated that assimilating inner-core data with GDAS covariance severely degrades the forecast for hurricanes.

Another important result is that the dropsonde data benefited all intensity categories (Figs. 21e–g), which represents a major improvement since T18. In particular, physics and DA improvements have apparently ameliorated the degradation seen when various types of reconnaissance data are assimilated in hurricanes. This reinforces the results from Zawislak et al. (2022), who showed that the combination of all reconnaissance data improved operational-HWRF intensity forecasts in a sample of landfalling U.S. hurricanes.

Nevertheless, dropsonde observations in this study did have intensity-dependent impacts. In general, the direct benefits of dropsondes occurred at earlier lead times for stronger TCs and at later lead times for weaker TCs. This is qualitatively similar to results obtained from an earlier reconnaissance impact assessment at NOAA that used the 2019 version of the operational HWRF, part of which was published in Zawislak et al. (2022). Stratifying the results shown in Fig. 4 of Zawislak et al. (2022) by TC intensity reveals that reconnaissance data improved VMAX forecasts of TS after about 48 h and of H345 before about 54 h (not shown). This general result was also true for various individual reconnaissance observing systems (e.g., TDR and dropsondes), though those results were never published. Yet, it is not clear whether the intensity-dependent impacts of reconnaissance data represent a general result or are specific to HWRF. For example, Sippel et al. (2022) found that adding reconnaissance data more robustly improved the forecasts of hurricanes than TS at longer lead times when using the operational GFS version 16.

[10] Note that dropsondes directly improved EPAC TC intensity and significant wind radii forecasts at short lead times and indirectly improved EPAC TC track and intensity forecasts at long lead times (see section 4). Findings were qualitatively similar to those for NATL TC, indicating that further forecast improvements are possible by sampling more EPAC TCs with dropsondes. However, the sample size for the EPAC was relatively small.

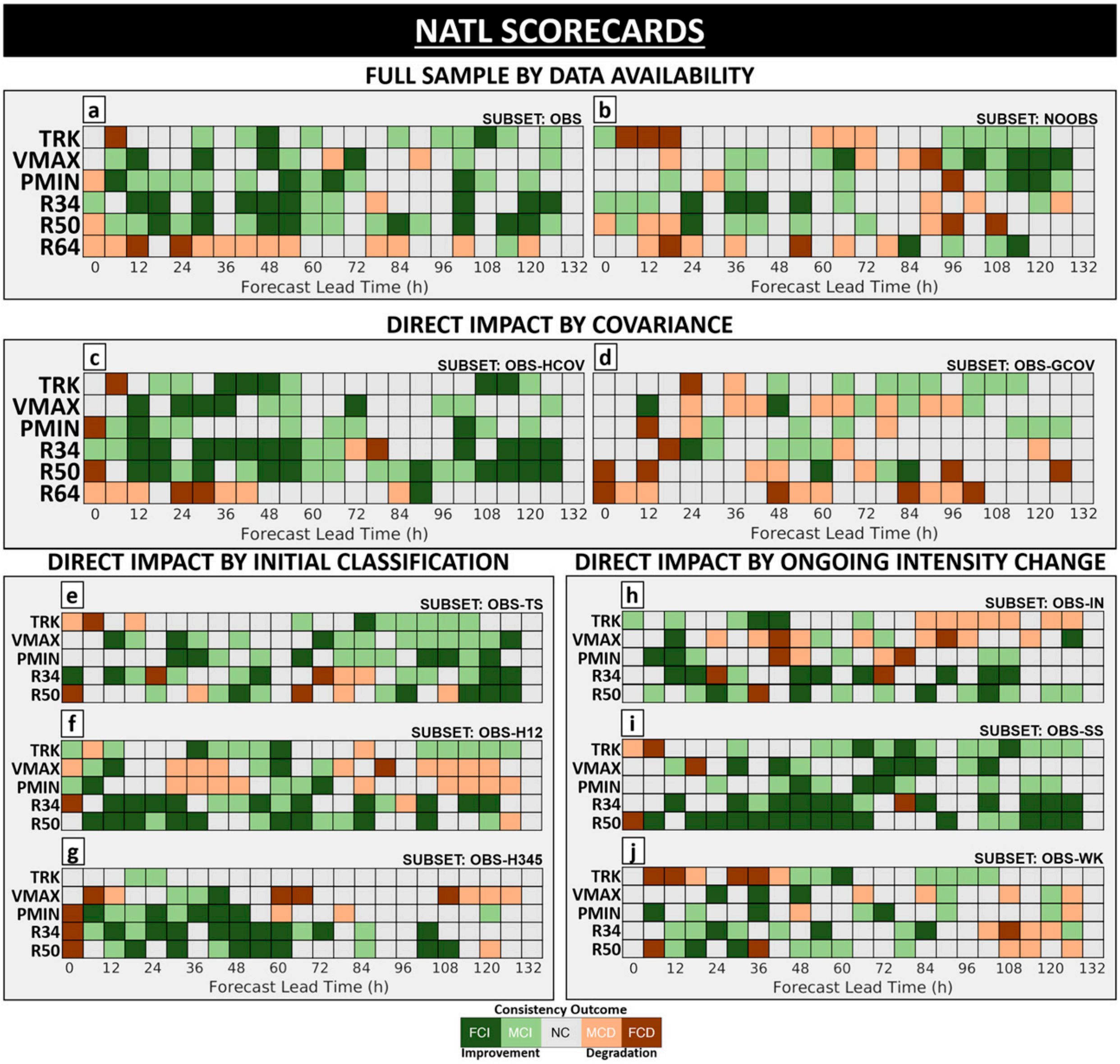


FIG. 21. Summary graphics of consistency metrics previously displayed for (a),(b) the full sample by data availability from Figs. 7, 11, and 15; (c),(d) the direct impact by covariance from Figs. 8, 12, and 16; (e)–(g) the direct impact by initial classification from Figs. 9a, 13a, and 17a; and (h)–(j) the direct impact by ongoing intensity change from Figs. 9b, 13b, and 17b.

The intensity-dependent results in Figs. 21e–g are particularly important when a TC is nearing landfall, as they indicate that expanding the frequency of sampling of hurricanes with dropsondes would lead to more accurate track, PMIN, and outer-wind-radii forecasts at short lead times. As described above, this would lead to more refined and effective watches and warnings as well as aid emergency managers and local officials. Improvements to PMIN forecasts in OBS-H345 between 6 and 48 h is particularly notable, since PMIN is better correlated with damage than VMAX (e.g., Chavas et al. 2017; Klotzbach et al. 2020) and is easier to accurately diagnose with dropsondes than TC-size.

Despite the recent improvements in HWRF, it appears that DA deficiencies still exist. For example, while assimilated dropsonde observations mostly improved steady-state (SS) TCs in this study (Fig. 21i), they had diminished benefits for TCs undergoing intensity change (cf. Figs. 21h,j). This is particularly true for VMAX, for which dropsonde observations improved forecasts 10%–20% in SS TCs but not at all when intensity was changing (Fig. 13b). This result could reflect the fact that even with HWRF-cycled covariance, not considering the evolution of covariance over the 6-h DA windows could cause large analysis errors in a changing TC. Indeed, Lu et al. (2017b) and Davis et al. (2021) showed using HWRF-cycled covariance that considered the time evolution within a 6-h window (through either more frequent DA cycling or 4DEnVAR) produces superior analyses and forecasts. Further, Christophersen et al. (2018) used 30-min cycling to assimilate dropsondes from NASA's high-altitude Global Hawk and obtained their best results for non-SS TCs. The present and previous results combined suggest that dropsondes and other inner-core data can have a greater impact on forecasts with further DA improvements.

Though results here suggest that with the appropriate DA treatment dropsondes can considerably improve many aspects of TC forecasts, a limitation difficult to overcome is that relatively few TCs have reconnaissance data. In this 2017–20 sample, only 30% of individual forecasts from NATL TCs have dropsondes. While assimilating dropsonde observations in every cycle of a TC would lead to even further forecast improvements, that is likely not feasible due to limited resources. Therefore, three suggestions on how to achieve additional forecast improvements are presented below.

First, the number of individual forecasts with assimilated dropsonde observations should be increased. Once a TC is forecast to be a landfall threat, it would be beneficial to assimilate dropsonde observations throughout the entire TC (i.e., no radial gaps) in every model cycle through landfall. This can be achieved with greater investment in reconnaissance resources to allow for expanded use of these observing systems. In addition to data from dropsondes, NCEP currently assimilates other reconnaissance data that are also known to improve forecasts, so increasing the amount of TC reconnaissance could carry forecast benefits far greater than those realized from dropsondes alone. Similar benefits would likely be extended to forecasts of pre-TC disturbances, though those cases were not evaluated in this study.

Second, improvements to dropsonde-observation processing are needed. Currently, models at NCEP do not use dropsonde data to their fullest capacity. They instead rely on postprocessed data from the mandatory and significant levels in WMO TEMP DROP messages. This is true even for the improved dropsonde treatment in HWRF, as described in section 2b. Ongoing work at AOML suggests that assimilating superobservations of full-resolution data should considerably benefit analyses and forecasts (Sellwood et al. 2020).

Finally, investing in the development of superior DA methods would also likely improve the impact of dropsondes as well as all other inner-core data. This paper demonstrated that dropsondes generally only benefited forecasts when HWRF-cycled covariance was used for DA. Results suggest that further improvements are likely with DA methods that consider the time-evolution of covariance over shorter periods. Other opportunities to improve inner-core DA include tuning the DA system with improved accounting of observation errors and quality control (e.g., Aksoy et al. 2022) and improving the manner in which observations can simultaneously impact the analysis at multiple scales (e.g., Zhang et al. 2009; Huang et al. 2021). Further improvements to inner-core data impact can likely be achieved by using cutting-edge techniques designed to handle non-Gaussian error distributions, such as particle filters (e.g., Kurosawa and Poterjoy 2022; Poterjoy 2022). For more details on these potential improvements, see Christophersen et al. (2022). Given present-day suboptimalities in both observation processing and DA, the results here likely represent a floor for dropsonde impact upon which much greater advancements can be achieved.

Results presented in this study addressed the overall impact of dropsondes on TC forecasts. While comprehensive, one of the questions left unanswered is how different dropsonde sampling strategies might impact forecasts. A companion study to the present one is quantifying the impact of dropsonde sampling as a function of radius from the TC center. Additionally, as described in section 1, some studies have assessed the impact of dropsondes from specific aircraft and the impact of dropsondes that target specific regions near the TC or in the synoptic environment. Two other large-sample, multiyear studies are under way to specifically examine the impact of various flight-track patterns on TC forecasts. Results found from all of these studies will help optimize dropsonde sampling during reconnaissance missions.

Acknowledgments. This research was carried out in part under the auspices of the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Cooperative Institute of the University of Miami and the National Oceanic and Atmospheric Administration, Cooperative Agreement NA20OAR4320472 while the lead author (Sarah Ditchek) was supported by the FY18 Hurricane Supplemental (NOAA Award NA19OAR0220188). Thank you to Andrew Kren and Karina Apodaca for their guidance and administrative assistance, AOML/HRD's Xuejin Zhang and Sundararaman (Gopal) Gopalakrishnan for allocating HPC resources for the experiments, Michael Brennan, Kelly Ryan, and three WAF reviewers (Hui Christophersen, Zhan Zhang, and one anonymous) for their constructive comments on the manuscript, Ron McTaggart-Cowan for his suggestions on how to condense the manuscript so that case study examples could be added, and EMC and DTC (especially Evan Kalina, Zhan Zhang, Linlin Pan, Biju Thomas, and Mrinal Biswas) for their help with model issues encountered. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author(s) and do not necessarily reflect those of OAR or the Department of Commerce.

Data availability statement. Experiments were performed on the NOAA RDHPCS supercomputers Hera, Orion, and WCOSS, with output archived on NCEI's High Performance Storage System (HPSS) for a 5-yr term. This output is not publicly available, however, those interested in the output can contact the lead author. 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. Dropsonde data can be found on HRD's Hurricane Field Program's public-facing website at https://www.aoml.noaa.gov/data-products/. Finally, the Graphics for OS(S)Es and Other modeling applications on TCs (GROOT) verification package developed by the lead author and funded by the Quantitative Observing System Assessment

Program (QOSAP) and the FY18 Hurricane Supplemental (NOAA Award NA19OAR0220188) was used to generate graphics for this publication. It can be found at https://github.com/sditchek/GROOT.

REFERENCES

Aberson, S. D., 2002: Two years of operational hurricane synoptic surveillance. *Wea. Forecasting*, **17**, 1101–1110, https://doi.org/10.1175/1520-0434(2002)017<1101:TYOOHS>2.0.CO;2.

——, 2003: Targeted observations to improve operational tropical cyclone track forecast guidance. *Mon. Wea. Rev.*, **131**, 1613–1628, https://doi.org/10.1175//2550.1.

——, 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.

——, 2011: The impact of dropwindsonde data from the THORPEX Pacific Area Regional Campaign and the NOAA hurricane field program on tropical cyclone forecasts in the Global Forecast System. *Mon. Wea. Rev.*, **139**, 2689–2703, https://doi.org/10.1175/2011MWR3634.1.

——, and J. L. Franklin, 1999: Impact on hurricane track and intensity forecasts of GPS dropwindsonde observations from the first-season flights of the NOAA Gulfstream-IV jet aircraft. *Bull. Amer. Meteor. Soc.*, **80**, 421–428, https://doi.org/10.1175/1520-0477(1999)080<0421:IOHTAI>2.0.CO;2.

——, and B. J. Etherton, 2006: Targeting and data assimilation studies during Hurricane Humberto (2001). *J. Atmos. Sci.*, **63**, 175–186, https://doi.org/10.1175/JAS3594.1.

——, 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. J. Cione, B. A. 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.

——, ——, ——, Z. Zhang, F. D. Marks, and R. Atlas, 2019: Track uncertainty in high-resolution HWRF ensemble forecasts of Hurricane Joaquin. *Wea. Forecasting*, **34**, 1889–1908, https://doi.org/10.1175/WAF-D-19-0028.1.

——, D. Sheinin, B. Thomas, L. Gramer, Z. Zhang, B. Liu, H.-S. Kim, and A. Mehra, 2020: A hydrodynamical atmosphere/ocean coupled modeling system for multiple tropical cyclones. *Atmosphere*, **11**, 869, https://doi.org/10.3390/atmos11080869.

——, X. Zhang, and S. G. Gopalakrishnan, 2022: High-definition hurricanes: Improving forecasts with storm-following nests. *Bull. Amer. Meteor. Soc.*, **103**, E680–E703, https://doi.org/10.1175/BAMS-D-20-0134.1.

Biswas, M. K., and Coauthors, 2018: Hurricane Weather Research and Forecasting (HWRF) model: 2018 Scientific documentation. Scientific Doc. HWRF v4.0a, 112 pp., https://dtcenter.org/sites/default/files/community-code/hwrf/docs/scientific_documents/HWRFv4.0a_ScientificDoc.pdf.

Burpee, R. W., 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, 76 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf.

Chavas, D. R., K. A. Reed, and J. A. Knaff, 2017: Physical understanding of the tropical cyclone wind-pressure relationship. *Nat. Commun.*, **8**, 1360, https://doi.org/10.1038/s41467-017-01546-9.

Chou, K.-H., C.-C. Wu, P.-H. Lin, S. D. Aberson, M. Weissmann, F. Harnisch, and T. Nakazawa, 2011: The impact of dropwindsonde observations on typhoon track forecasts in DOTSTAR and T-PARC. *Mon. Wea. Rev.*, **139**, 1728–1743, https://doi.org/10.1175/2010MWR3582.1.

Christophersen, H., A. Aksoy, J. Dunion, and K. Sellwood, 2017: The impact of NASA Global Hawk unmanned aircraft dropwindsonde observations on tropical cyclone track, intensity, and structure: Case studies. *Mon. Wea. Rev.*, **145**, 1817–1830, https://doi.org/10.1175/MWR-D-16-0332.1.

——, ——, ——, and S. Aberson, 2018: Composite impact of Global Hawk unmanned aircraft dropwindsondes on tropical cyclone analyses and forecasts. *Mon. Wea. Rev.*, **146**, 2297–2314, https://doi.org/10.1175/MWR-D-17-0304.1.

——, J. Sippel, A. Aksoy, and N. L. Baker, 2022: Recent advancements for tropical cyclone data assimilation. *Ann. N. Y. Acad. Sci.*, **1517**, 25–43, https://doi.org/10.1111/nyas.14873.

Davis, B., X. Wang, and X. Lu, 2021: A comparison of HWRF six-hourly 4DEnVar and hourly 3DEnVar assimilation of inner core tail Doppler radar observations for the prediction of Hurricane Edouard (2014). *Atmosphere*, **12**, 942, https://doi.org/10.3390/atmos12080942.

Ditchek, S. D., J. A. Sippel, G. J. Alaka, S. B. Goldenberg, and L. Cucurull, 2022: A systematic assessment of dropsonde impact during the 2017–2020 hurricane seasons using the basin-scale HWRF: Overall impacts. *35th Conf. on Hurricanes and Tropical Meteorology*, New Orleans, LA, Amer. Meteor. Soc., 6B.2, https://ams.confex.com/ams/35Hurricanes/meetingapp.cgi/Paper/401287.

——, ——, P. Marinescu, and G. J. Alaka, 2023: Improving best track verification of tropical cyclones: A new metric to identify forecast consistency. *Wea. Forecasting*, **38**, 817–831, https://doi.org/10.1175/WAF-D-22-0168.1.

Franklin, J. L., and M. DeMaria, 1992: The impact of omega dropwindsonde observations on barotropic hurricane track forecasts. *Mon. Wea. Rev.*, **120**, 381–391, https://doi.org/10.1175/1520-0493(1992)120<0381:TIOODO>2.0.CO;2.

Goldenberg, S. B., S. G. Gopalakrishnan, V. Tallapragada, T. Quirino, F. Marks, S. Trahan, X. Zhang, and R. Atlas, 2015: The 2012 triply nested, high-resolution operational version of the Hurricane Weather Research and Forecasting Model (HWRF): Track and intensity forecast verifications. *Wea. Forecasting*, **30**, 710–729, https://doi.org/10.1175/WAF-D-14-00098.1.

Harnisch, F., and M. Weissmann, 2010: Sensitivity of typhoon forecasts to different subsets of targeted dropsonde observations. *Mon. Wea. Rev.*, **138**, 2664–2680, https://doi.org/10.1175/2010MWR3309.1.

Huang, B., X. Wang, D. T. Kleist, and T. Lei, 2021: A simultaneous multiscale data assimilation using scale-dependent localization in GSI-based hybrid 4dEnVar for NCEP FV3-based GFS. *Mon. Wea. Rev.*, **149**, 479–501, https://doi.org/10.1175/MWR-D-20-0166.1.

Klotzbach, P. J., M. M. Bell, S. G. Bowen, E. J. Gibney, K. R. Knapp, and C. J. Schreck, 2020: Surface pressure a more skillful predictor of normalized hurricane damage than maximum sustained wind. *Bull. Amer. Meteor. Soc.*, **101**, E830–E846, https://doi.org/10.1175/BAMS-D-19-0062.1.

Kren, A., L. Cucurull, and H. Wang, 2018: Impact of UAS Global Hawk dropsonde data on tropical and extratropical cyclone forecasts in 2016. *Wea. Forecasting*, **33**, 1121–1141, https://doi.org/10.1175/WAF-D-18-0029.1.

Kurosawa, K., and J. Poterjoy, 2022: A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble Kalman filters for data assimilation. *Mon. Wea. Rev.*, **151**, 105–125, https://doi.org/10.1175/MWR-D-22-0108.1.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Lu, X., X. Wang, Y. Li, M. Tong, and X. Ma, 2017a: GSI-based ensemble-variational hybrid data assimilation for HWRF for hurricane initialization and prediction: Impact of various error covariances for airborne radar observation assimilation. *Quart. J. Roy. Meteor. Soc.*, **143**, 223–239, https://doi.org/10.1002/qj.2914.

——, ——, M. Tong, and V. Tallapragada, 2017b: GSI-based, continuously cycled, dual-resolution hybrid ensemble–variational data assimilation system for HWRF: System description and experiments with Edouard (2014). *Mon. Wea. Rev.*, **145**, 4877–4898, https://doi.org/10.1175/MWR-D-17-0068.1.

Majumdar, S. J., 2016: A review of targeted observations. *Bull. Amer. Meteor. Soc.*, **97**, 2287–2303, https://doi.org/10.1175/BAMS-D-14-00259.1.

——, M. J. Brennan, and K. Howard, 2013: The impact of dropwindsonde and supplemental rawinsonde observations on track forecasts for Hurricane Irene (2011). *Wea. Forecasting*, **28**, 1385–1403, https://doi.org/10.1175/WAF-D-13-00018.1.

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.

——, 2021: Important factors in the tracking of tropical cyclones in operational models. *J. Appl. Meteor. Climatol.*, **60**, 1265–1284, https://doi.org/10.1175/JAMC-D-20-0175.1.

NOAA, 2020: National hurricane operations plan. Office of the Federal Coordinator for Meteorological Services and Supporting Research (OFCM) Doc. FCM-P12-2020, NOAA, 178 pp., https://www.icams-portal.gov/resources/ofcm/nhop/2020_nhop.pdf.

Poterjoy, J., 2022: Implications of multivariate non-Gaussian data assimilation for multiscale weather prediction. *Mon. Wea. Rev.*, **150**, 1475–1493, https://doi.org/10.1175/MWR-D-21-0228.1.

Powell, M. D., and T. A. Reinhold, 2007: Tropical cyclone destructive potential by integrated kinetic energy. *Bull. Amer. Meteor. Soc.*, **88**, 513–526, https://doi.org/10.1175/BAMS-88-4-513.

Pu, Z., X. Li, C. S. Velden, S. D. Aberson, and W. T. Liu, 2008: The impact of aircraft dropsonde and satellite wind data on numerical simulations of two landfalling tropical storms during the tropical cloud systems and processes experiment. *Wea. Forecasting*, **23**, 62–79, https://doi.org/10.1175/2007WAF2007006.1.

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.

Sellwood, K., J. A. Sippel, and A. Aksoy, 2020: Optimizing dropwindsonde levels for data assimilation. *20th Symp. on Meteorological Observation and Instrumentation*, Boston, MA, Amer. Meteor. Soc., 5.6, https://ams.confex.com/ams/2020Annual/webprogram/Paper365847.html.

Shi, J. J., S. Chang, and S. Raman, 1996: Impact of assimilations of dropwindsonde data and SSM/I rain rates on numerical predictions of Hurricane Florence (1988). *Mon. Wea. Rev.*, **124**, 1435–1448, https://doi.org/10.1175/1520-0493(1996)124<1435:IOAODD>2.0.CO;2.

Simpson, R. H., and H. Saffir, 1974: The hurricane disaster potential scale. *Weatherwise*, **27**, 169–186.

Sippel, J. A., 2020: The use of reconnaissance aircraft data in weather forecast models. NOAA (SECART) 2020 Hurricane Awareness Webinar Series, NOAA, accessed 27 May 2020, https://www.noaa.gov/regions/2020-hurricane-awareness-webinars.

——, Z. Zhang, L. Bi, and A. Mehra, 2021: Recent advances in operational HWRF data assimilation. *34th Conf. on Hurricanes and Tropical Meteorology*, online, Amer. Meteor. Soc., 3C.2, https://ams.confex.com/ams/34HURR/meetingapp.cgi/Paper/372789.

——, X. Wu, S. D. Ditchek, V. Tallapragada, and D. T. Kleist, 2022: Impacts of assimilating additional reconnaissance data on operational GFS tropical cyclone forecasts. *Wea. Forecasting*, **37**, 1615–1639, https://doi.org/10.1175/WAF-D-22-0058.1.

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.

Torn, R., 2020: Transitioning ensemble-based TC track and intensity sensitivity to operations: Current status and future plans. *Tropical Cyclone Operations and Research Forum: Joint Hurricane Testbed JHT*, online, NOAA, 17 pp., https://www.nhc.noaa.gov/jht/19-22reports/JHT1922_IHC_2020_Torn.pdf.

——, 2021: Transitioning ensemble-based TC track and intensity sensitivity to operations: Current status and future plans. *Tropical Cyclone Operations and Research Forum, Joint Hurricane Testbed JHT*, online, NOAA, 14 pp., https://www.nhc.noaa.gov/jht/19-22reports/JHT1922_IHC_2021_Torn.pdf.

——, and C. Snyder, 2012: Uncertainty of tropical cyclone best-track information. *Wea. Forecasting*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

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

Velden, C., and S. Goldenberg, 1987: The inclusion of high density satellite wind information in a barotropic hurricane-track forecast model. Preprints, *17th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 93 pp.

Weng, Y., and F. Zhang, 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.

Wu, C.-C., J.-H. Chen, P.-H. Lin, and K.-H. Chou, 2007: Targeted observations of tropical cyclone movement based on the adjoint-derived sensitivity steering vector. *J. Atmos. Sci.*, **64**, 2611–2626, https://doi.org/10.1175/JAS3974.1.

——, S.-G. Chen, C.-C. Yang, P.-H. Lin, and S. D. Aberson, 2012: Potential vorticity diagnosis of the factors affecting the track of Typhoon Sinlaku (2008) and the impact from dropwindsonde data during T-PARC. *Mon. Wea. Rev.*, **140**, 2670–2688, https://doi.org/10.1175/MWR-D-11-00229.1.

Yamaguchi, M., T. Iriguchi, T. Nakazawa, and C.-C. Wu, 2009: An observing system experiment for Typhoon Conson (2004) using a singular vector method and DOTSTAR data. *Mon. Wea. Rev.*, **137**, 2801–2816, https://doi.org/10.1175/2009MWR2683.1.

Zawislak, J., and Coauthors, 2022: 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. A. Sippel, Z. Meng, and C. H. Bishop, 2009: Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2105–2125, https://doi.org/10.1175/2009MWR2645.1.

Zhang, X., S. G. Gopalakrishnan, S. Trahan, T. S. Quirino, Q. Liu, Z. Zhang, G. Alaka, and V. Tallapragada, 2016: Representing multiple scales in the Hurricane Weather Research and Forecasting modeling system: Design of multiple sets of movable multilevel nesting and the basin-scale HWRF forecast application. *Wea. Forecasting*, **31**, 2019–2034, https://doi.org/10.1175/WAF-D-16-0087.1.

Zhang, Z., J. A. Zhang, G. J. Alaka Jr., K. Wu, A. Mehra, and V. Tallapragada, 2021: A statistical analysis of high-frequency track and intensity forecasts from NOAA's operational Hurricane Weather Research and Forecasting (HWRF) modeling system. *Mon. Wea. Rev.*, **149**, 3325–3339, https://doi.org/10.1175/MWR-D-21-0021.1.