A Comparison of the Impacts of Inner-Core, In-Vortex, and Environmental Dropsondes on Tropical Cyclone Forecasts during the 2017–20 Hurricane Seasons

SARAH D. DITCHEK[a,b] AND JASON A. SIPPEL[b]

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

(Manuscript received 22 March 2023, in final form 17 August 2023, accepted 21 August 2023)

ABSTRACT: This study conducts the first large-sample comparison of the impact of dropsondes in the tropical cyclone (TC) inner core, vortex, and environment on NWP-model TC forecasts. We analyze six observing-system experiments, focusing on four sensitivity experiments that denied dropsonde observations within annuli corresponding with natural breakpoints in reconnaissance sampling. These are evaluated against two other experiments detailed in a recent parallel study: one that assimilated and another that denied dropsonde observations. Experiments used a basin-scale, multistorm configuration of the Hurricane Weather Research and Forecasting (HWRF) Model and covered active periods of the 2017–20 North Atlantic hurricane seasons. Analysis focused on forecasts initialized with dropsondes that used mesoscale error covariance derived from a cycled HWRF ensemble, as these forecasts were where dropsondes had the greatest benefits in the parallel study. Some results generally support findings of previous research, while others are novel. Most notable was that removing dropsondes anywhere, particularly from the vortex, substantially degraded forecasts of maximum sustained winds. Removing in-vortex dropsondes also degraded outer-wind-radii forecasts in many instances. As such, in-vortex dropsondes contribute to a majority of the overall impacts of the dropsonde observing system. Additionally, track forecasts of weak TCs benefited more from environmental sampling, while track forecasts of strong TCs benefited more from in-vortex sampling. Finally, inner-core-only sampling strategies should be avoided, supporting a change made to the U.S. Air Force Reserve's sampling strategy in 2018 that added dropsondes outside of the inner core.

SIGNIFICANCE STATEMENT: This study uses a regional hurricane model to conduct the most comprehensive assessment to date of the impact of dropsondes at different distances away from the center of a tropical cyclone (TC) on TC forecasts. The main finding is that in-vortex dropsondes are most important for intensity and outer-wind-radii forecasts. Particularly notable is the impact of dropsondes on TC maximum wind speed forecasts, as reducing sampling anywhere would degrade those forecasts.

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

1. Introduction

Airborne reconnaissance conducted by the U.S. Air Force Reserve (USAF) and the National Oceanographic and Atmospheric Administration (NOAA) has improved NWP-model forecasts of tropical cyclones (TCs) for about forty years. The effort to use the data for NWP began in the early 1980s, when airborne missions began transmitting a limited amount of Global Positioning Satellite (GPS) dropsonde data for operational use (e.g., Burpee et al. 1984). The benefits of reconnaissance data have proven so great that they have led NOAA to procure aircraft (e.g., Rappaport et al. 2009) and to invest a great deal into improving real-time transmission and data assimilation (Zawislak et al. 2022). Given these developments, NHC has increasingly relied on NOAA aircraft for TC reconnaissance.

Numerous peer-reviewed studies have examined the impact of reconnaissance data on NWP. Since dropsondes have the longest history of operational assimilation among reconnaissance data types, all the earliest research naturally focused on how dropsonde data affected TC forecasts. Initial studies focused specifically on the impact of environmental dropsonde data on TC track forecasts (e.g., Burpee et al. 1996; Aberson and Franklin 1999). Later studies began assessing the impacts of dropsondes on both track and intensity forecasts, though the focus remained mostly on impacts of environmental dropsonde data (e.g., Aberson 2002, 2010, 2011). In particular, Aberson (2010, 2011) used fairly large samples to show that assimilating dropsonde data improves TC forecasts, a result that has been verified in a number of subsequent studies (see Ditchek et al. 2023a, hereafter D23A, their Fig. 1).

Starting around 2010, research began to focus on the impacts of assimilating more types of reconnaissance data, including that from dropsondes, from within the TC vortex. From these efforts, a number of studies found that assimilating high-resolution, in-vortex data considerably improves both track and intensity forecasts (e.g., Zhang et al. 2009, 2011; Weng and Zhang 2012; Aberson et al. 2015; Weng and Zhang 2016). Further, improved assimilation of inner-core data at the National Centers for Environmental Prediction (NCEP) has contributed to recent improvements in NWP intensity forecasts (Zawislak et al. 2022).

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

DOI: 10.1175/WAF-D-23-0055.1

Though we now know that in-vortex and environmental reconnaissance data both benefit TC forecasts, so far only Harnisch and Weissmann (2010) assessed their impacts relative to one another. They found that dropsondes in remote regions (i.e., 700–1200 km from the TC center) had less of an impact on track forecasts than did dropsondes in the vicinity of TCs (i.e., those between the core and remote regions). Despite their small sample (<20 cases), their results in part led NHC to focus high-altitude synoptic-surveillance missions closer to the TC center beginning with Hurricane Florence in 2018 (see D23A, their Fig. 3c).

To fill a gap in prior research, this study more comprehensively evaluates the varying impacts of dropsonde data in the TC inner core, broader vortex, and environment. To do so, an experimental version of the Hurricane Weather Research and Forecasting (HWRF) Model was used to conduct four observing-system sensitivity experiments that denied dropsondes at various annuli that correspond with natural breakpoints in reconnaissance sampling. Experiments were run over active periods in the North Atlantic basin (NATL; including the North Atlantic Ocean, the Gulf of Mexico, and the Caribbean Sea) during the 2017–20 hurricane seasons. Note that the experimental design used here (i.e., model, dropsonde data, specific date ranges used, TCs included, and most of the verification performed) mirrors that of D23A, which assessed the overall impacts of dropsondes on TC forecasts. Thus, this investigation not only serves as an extension of D23A, but it also enables comparisons between the four sensitivity experiments and the two D23A experiments (one that assimilated and another that denied dropsonde observations).

The remainder of this manuscript is structured as follows. Given the similarities with D23A, section 2 will provide only a brief overview of the experimental design. For a more detailed description, please see D23A. Section 2 then provides details on the four sensitivity experiments and additional verification techniques used. Finally, sections 3 and 4 detail results from the experiments while section 5 provides a summary and recommendations for future work.

2. Data and methods

a. Model

The sensitivity experiments presented here were conducted using the 2020 version of the experimental, parallel, basin-scale, multistorm version of HWRF (HB20; Zhang et al. 2016; Alaka et al. 2017; Alaka 2019; Alaka et al. 2020, 2022) 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). The parent domain (D01) spans the entire NATL and eastern North Pacific basins with a horizontal grid spacing of 13.5 km. HB20 is configured with movable nests (D02 and D03) for up to five TCs. D02 and D03 have cloud-permitting resolutions of 4.5 and 1.5 km, respectively, and are two-way interactive with each other and with D01 to transfer critical information about the TC and its environment across scales. Consequently, forecasts of all TCs in the basin were impacted if dropsonde observations were assimilated in any TC.

b. Experiment suite and scope

Dropsondes included in this assessment were the same as those used in D23A—those launched into TCs between 2017 and 2020 from four different types of aircraft: The U.S. Air Force Reserve's (USAF) low-mid altitude WC-130J (C-130), NOAA's low-mid altitude WP-3D (P-3), NOAA's high-altitude Gulfstream IV-SP (G-IV), and NASA's high-altitude Global Hawk (GH).[1] The resulting atmospheric profiles of quality-controlled pressure, temperature, relative humidity, and horizontal winds were transmitted in TEMP DROP messages from the aircraft (NOAA 2020), adjusted to account for the advection of dropsondes while falling (Aberson 2008; Aberson et al. 2017), and converted to the standard NCEP Prepared Binary Universal Form for the Representation of meteorological data (PrepBUFR) format before being assimilated into HB20.

On average, reconnaissance missions concentrate dropsonde sampling into three distinct annuli (Fig. 1). Inner-core dropsondes, which sample the eye and eyewall, typically fall within 75 km of the TC center. A natural minimum in dropsondes occurs at 75 km, with the frequency gradually increasing again until about 250 km. Those dropsondes between 75 and 250 km generally fall outside the inner core but still within the vortex. They mostly coincide with mid- and end-points of radial legs in the various low- and midaltitude reconnaissance patterns. Occasionally, a high-altitude circumnavigation at a radius of about 165 km also contributes to dropsonde sampling within this annulus. Another natural minimum occurs at 250 km, outside of which, high-altitude, environmental sampling accounts for the vast majority of dropsondes. See D23A for more information regarding these reconnaissance patterns.

As mentioned in the introduction, this study takes advantage of the natural breakpoints in sampling (Fig. 1b) to conduct four sensitivity experiments. The experiments denied dropsonde data either outside (Figs. 2b,c) or inside (Figs. 2e,f) the 75- and 250-km radii, respectively, and are named as follows: No Inner Core (NO-IC; Fig. 2b), No Vortex (NO-VOR; Fig. 2c), Inner Core Only (ICO); Fig. 2e, and No Environment (NO-ENV; Fig. 2f). Note that *individual dropsonde observations* rather than *entire dropsonde profiles* were denied in these sensitivity experiments. As such, some individual dropsondes could yield data within two adjacent annuli due to the advection of dropsondes while falling.

These four experiments were compared to ALL and NO discussed in D23A. Their ALL (here called ALL-DROP for clarity; Fig. 2a) assimilated dropsonde observations, whereas their NO (here called NO-DROP; Fig. 2d) did not. Therefore, these sensitivity experiments covered the same periods described in D23A (their Table 1) and included 634 cycles, resulting in 2139 individual forecasts. These forecasts covered 92 TCs, 41 of which had assimilated dropsonde observations. Note that other than modifying how dropsonde observations

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

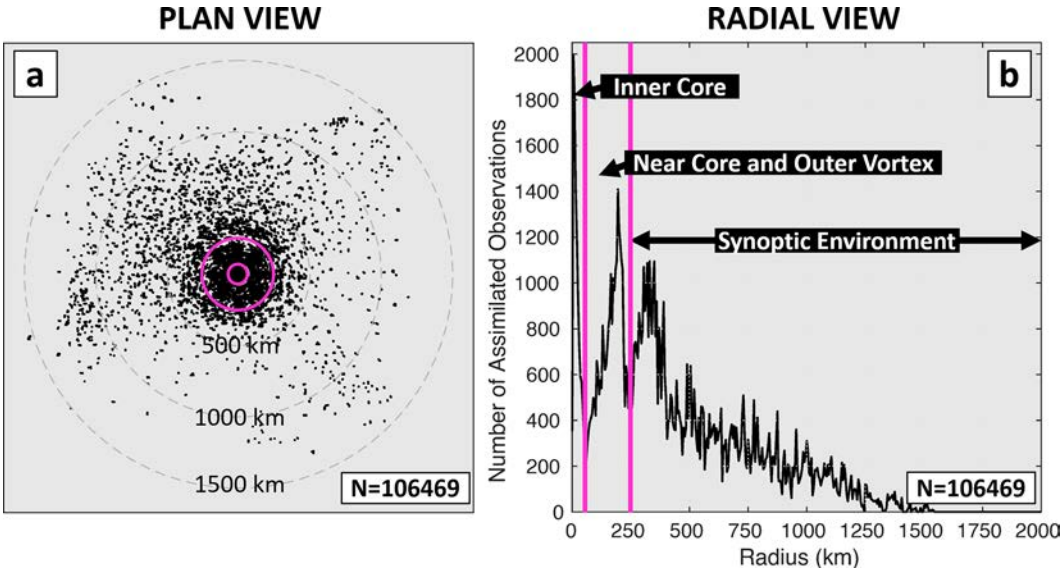

FIG. 1. The number of individually assimilated dropsonde temperature observations in each TC's D02 for the full sample in D23A in (a) plan view and (b) radial view. Observations are shown in a TC-relative framework, where the location of the TC center is interpolated to the time of each individual dropsonde observation before calculating the azimuth and radial location of the observation relative to the TC center. Note that dropsonde humidity and wind observations were also assimilated, though are not shown for simplicity. Pink lines indicate the two breakpoints chosen to design the four experiments: 75 and 250 km. This figure is identical to Figs. 4a and 4b in D23A except for the annotations in (b) and the overlaying pink circles and lines at two radii: 75 and 250 km.

were assimilated, all experiments otherwise assimilated all conventional, reconnaissance, and satellite data assimilated into HB20.

For reference, Fig. 3 shows the TC-relative plan-view and radial-view distributions of the number of individually assimilated dropsonde temperature observations in D02 for each of the four experiments. Note that for ICO and NO-ENV, some dropsondes are present well outside of 75 and 250 km, respectively. This only occurred when the D02 of two different TCs sufficiently overlapped that the domains covered at least part of both TCs. For example, consider ICO, which denied dropsonde observations outside of a 75-km radius of the TC center. Observations within 75 km in one TC could also be assimilated in the outer regions of a second TC. Since the same dropsonde could be assimilated into multiple TCs with overlapping D02, data points in Fig. 3 do not correspond 1-to-1 with actual dropsonde data, especially at radii > 1000 km.

As will become clear in the results, it is useful to understand how dropsonde sampling varies as a function of TC intensity as well as how well dropsondes in the various experiments sample the outer wind radii. As such, Fig. 4 depicts 1) the distribution of dropsonde observations overall and by initial classification (i.e., columns 1–2) and 2) histograms of best track outer wind radii values (i.e., R34 and R50; column 3). Each initial classification had a symmetric distribution of coverage within 250 km, but between 250 and 500 km there was a slight northward sampling bias that decreased for stronger TCs (Fig. 4, column 1). While a similar number of observations were assimilated in each initial classification, differences in radial distribution did occur (Fig. 4, column 2). For example, the stronger the TC, the more the inner core and environment was sampled. Interestingly, the peak in H345 sampling occurred outside of 250 km, whereas for H12 the peak was inside that radius. Finally, note that the entire R34 and R50 distributions shift radially outward as TCs get stronger, particularly when comparing tropical storms to hurricanes (Fig. 4, column 3). This has some relevance to R34 and R50 results, so these graphics will be referred back to later in the text.

c. *Verification*

As mentioned in the introduction, this study follows the framework of D23A. A main finding in D23A was that sampling TCs with dropsondes can directly improve TC forecasts only if using sufficiently advanced data assimilation (DA) techniques (i.e., HWRF-cycled mesoscale error covariance; their Fig. 21c). In contrast, dropsondes had neutral impacts on the forecast when using global-model error covariance during DA (their Fig. 21d). These results reinforced the fact that appropriate DA treatment is crucial for improved TC forecasts (e.g., Zhang et al. 2009; Lu et al. 2017; Tong et al. 2018). Thus, this paper focuses on those forecasts with direct sampling by dropsondes that also utilized HWRF-cycled mesoscale error covariance, which D23A referred to as OBS-HCOV. Note that we treat OBS-HCOV from D23A as our "full sample" and call it such, for brevity.

The performance of each experiment was evaluated only for NATL TCs by verifying forecasts against the NHC "best track" (Landsea and Franklin 2013) available from NHC following the standard NHC Forecast Verification procedures (Cangialosi 2022). Note that while uncertainties in position, intensity, and

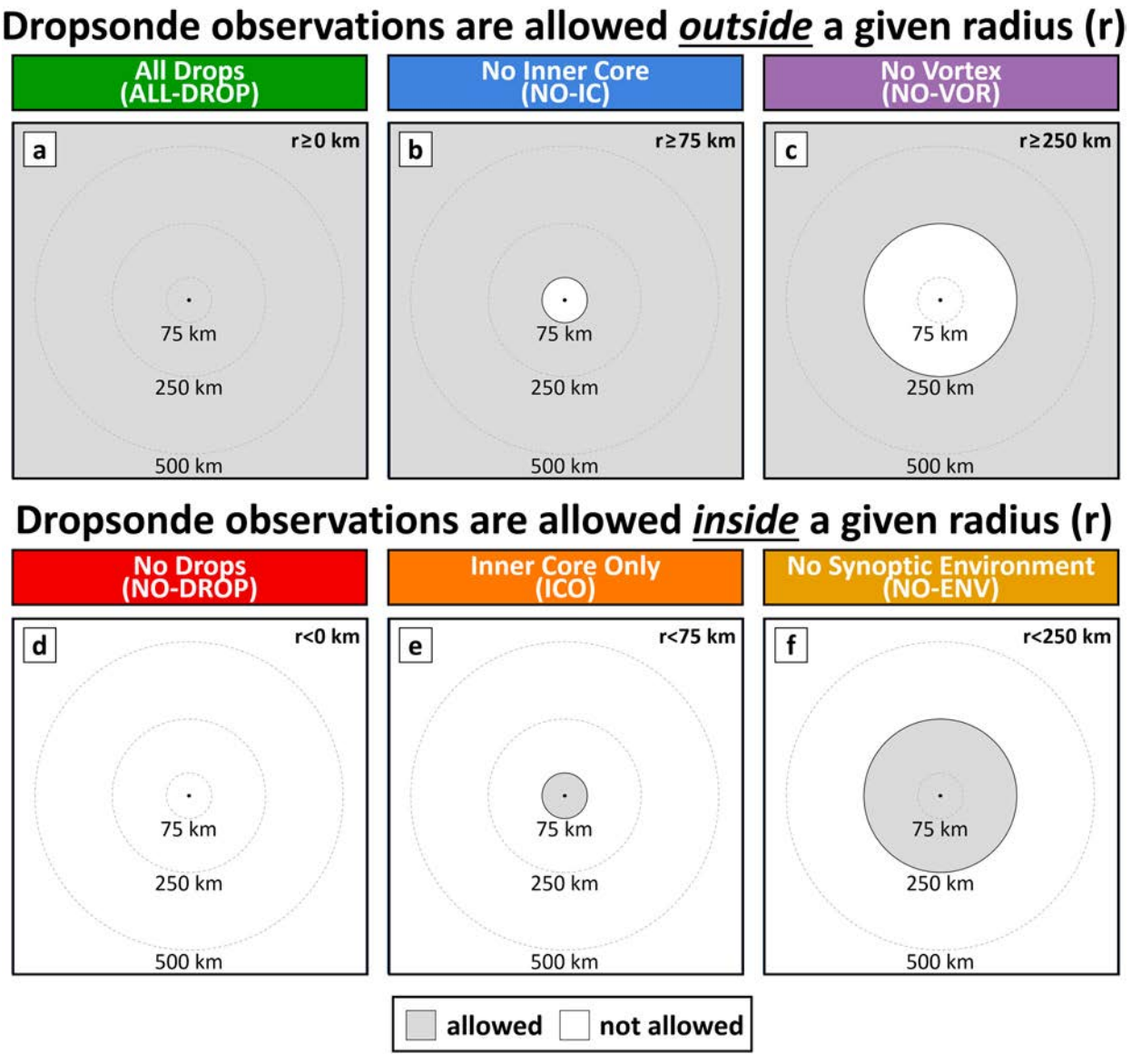

FIG. 2. Plan-view schematics of (a),(d) the two experiments from D23A and (b),(c),(e),(f) the four sensitivity experiments. Light gray shading indicates where dropsonde observations were allowed to be assimilated. The corresponding radii of the gray shading are given in the top right of each subplot. The colors of each experiment used in this figure will be used throughout the paper.

significant wind radii are present in the best track (Torn and Snyder 2012; Landsea and Franklin 2013) and in tracker output (Zhang et al. 2021), these are not taken into account in current TC-verification techniques. Further, results in this paper had no additional postprocessing (e.g., interpolation to produce "early" model forecasts; Cangialosi 2022). Thus, results are from the raw output (i.e., late-model results) from the Geophysical Fluid Dynamics Laboratory (GFDL) vortex tracker (Marchok 2002, 2021).

The choice to focus exclusively on late-model results is motivated by a few factors. First, recent operational testing with HWRF has shown that changing the details of early model interpolation can change maximum sustained 10-m wind speed MAE skill by as much as 2%–3% over multiyear, full-basin samples (not shown). Additionally, there is no unambiguous "best practice" for applying the interpolator, as different choices improve the early forecast at some lead times while degrading it at others. Finally, the optimal interpolator application also depends strongly on the TC characteristics and on the amount of inner-core data gathered. As such, the ideal early model interpolator configuration for the sample examined in this study would likely vary from the ideal configuration in operations. Since the goal of this study is to examine the impact of sampling differences on the subsequent forecast, we feel that nuances in application of the early model interpolation would likely obfuscate results.

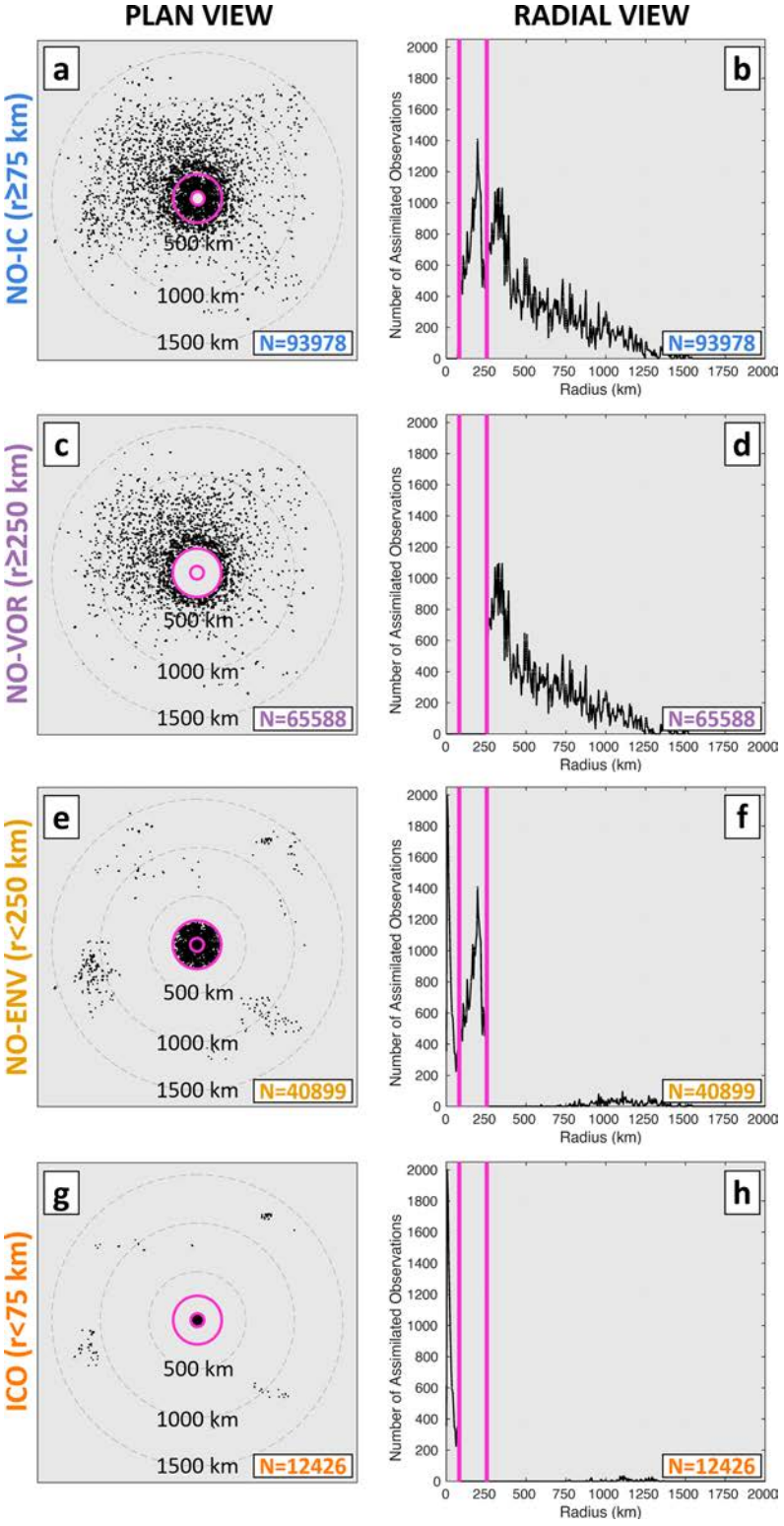

FIG. 3. The number of individually assimilated dropsonde temperature observations in each TC's D02 in (a),(b) NO-IC; (c),(d) NO-VOR; (e),(f) NO-ENV, and (g),(h) ICO in TC-relative (left) plan view and (right) radial view. Pink lines indicate the two divisors chosen to design the four experiments: 75 and 250 km.

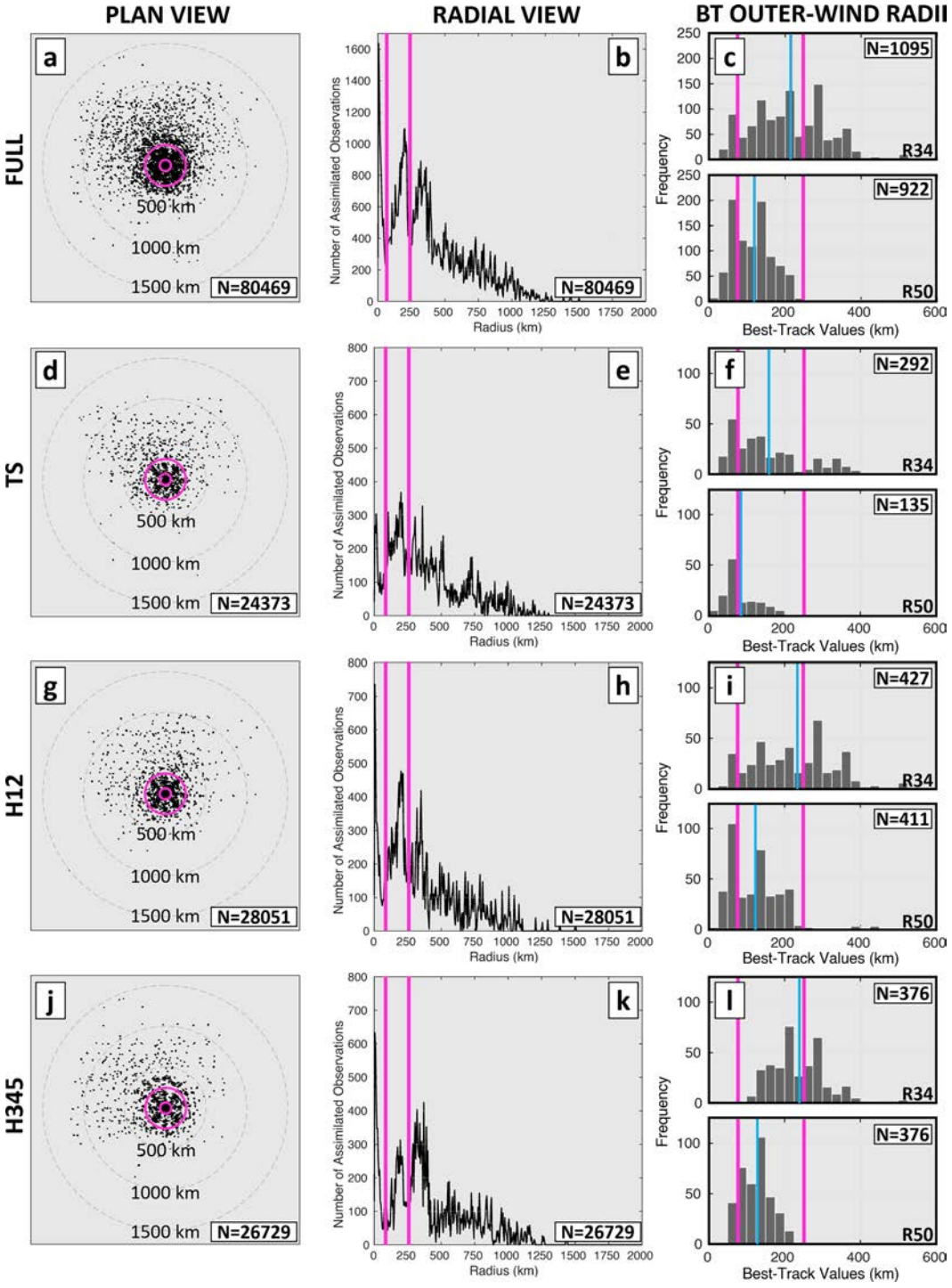

FIG. 4. The number of individually assimilated dropsonde temperature observations in each TC's D02 in ALL-DROP in TC-relative (left) plan view and (center) radial view as well as (right) histograms of the best track (BT) outer-wind-radii values (a)–(c) overall and stratified by (d)–(f) TS, (g)–(i) H12, and (j)–(l) H345. Pink lines indicate the two breakpoints chosen to design the four experiments: 75 and 250 km. Blue lines in the right column indicate the mean BT outer-wind radii value for FULL, TS, H12, and H345: 216, 159, 234, and 239 for R34, respectively, and 120, 84, 124, and 128 for R50, respectively.

Variables assessed include track, two measures of TC intensity [maximum sustained 10-m wind speed (VMAX); minimum mean sea level pressure (PMIN)], as well as the two outer surface-wind radii[2] reported by NHC [34-kt wind radii (R34) and 50-kt wind radii (R50); 1 kt ≈ 0.51 m s^{-1}]. Note that for R34 and R50, all quadrants were included individually in each sample. Finally, all results presented in this paper are for homogeneous samples. To be included in the homogeneous sample: 1) for a given cycle, all experiments had to satisfy the standard NHC forecast verification procedures and 2) a nonzero numeric value had to exist for a given variable in all experiments. Note that this second condition only impacted R34 and R50 samples. Both of these conditions are the reason for sample-size discrepancies between D23A and full-sample results presented in this paper of ≥96 h for track, VMAX, and PMIN as well as at all lead times for R34 and R50.

This study also explores how the sensitivity to changes in observing-system sampling evolves with the TC life cycle. To do so, it stratifies the full sample by initial classification into four groups according to their best track Saffir–Simpson scale (Simpson and Saffir 1974) 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}). As in D23A, TDs rarely had assimilated dropsonde observations and will not be included due to their small sample size.

Since land impacts can obfuscate interpretation of results (e.g., errors in landfall timing can dramatically increase VMAX errors, but landfall will ultimately diminish differences between experiments by driving all forecasts to a similar weak intensity), we also examine verification over water only for both the full sample (FULL-W) and initial classification stratifications (e.g., TS-W). More specifically, the overwater sample both 1) excludes an entire forecast if the TC was over land at the initial time and 2) excludes specific forecast lead times when the TC was initially over water but forecasted to later be over land. This procedure carries a caveat that it reduces the sample size at all lead times, and at long lead times the overwater sample becomes undesirably small. As will become apparent, this limitation shapes some of the approach to how results are presented. Note that discussion of results will generally focus on areas of agreement between the full and overwater samples.

As in D23A, this paper makes use of the consistency metric introduced in (Ditchek et al. 2023b, hereafter D23B) to objectively identify forecast lead times with fully consistent or marginally consistent improvement or degradation. This metric effectively identifies those lead times with consistency by requiring that the mean absolute error (MAE) skill, median absolute error (MDAE) skill, and frequency of superior performance (FSP; Velden and Goldenberg 1987; Goldenberg et al. 2015) exceed specified thresholds.

While the consistency metric alone is useful to compare ALL-DROP and NO-DROP, this study examines consistency at a somewhat higher level than in D23A in order to facilitate comparison of the multitude of experiments over various lead times. To do so, this study builds on the consistency-score methodology described in D23B to objectively indicate the impact of the observing-system changes relative to ALL-DROP. Briefly, the consistency-score methodology assigns numeric values of 1 (fully consistent improvement), 0.5 (marginally consistent improvement), 0 (no consistency), −0.5 (marginally consistent degradation), and −1 (fully consistent degradation) to the outcomes of the consistency analysis (e.g., Fig. 5a). These values are then averaged across either all lead times or a segment of lead times for each experiment and multiplied by 100 to generate a score that can be used to rank experiments by their overall consistency (e.g., Fig. 5b). This score ranges from −100% (which indicates that all lead times considered had fully consistent degradation) to 100% (which indicates that all lead times considered had fully consistent improvement).

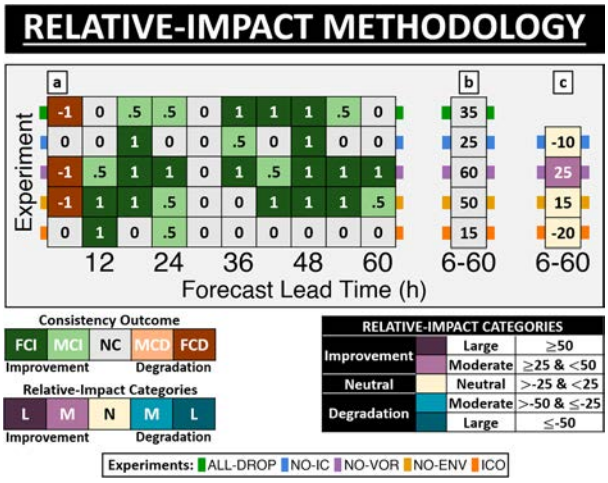

FIG. 5. A demonstration of the relative-impact methodology. Colored rectangles at the left and right edges of each panel indicate the experiment: ALL-DROP (green), NO-IC (blue), NO-VOR (purple), NO-ENV (yellow), and ICO (orange). (a) The consistency-metric scorecard for each experiment relative to NO-DROP at short lead times (6–60 h) colored by the consistency-outcome scale where "FC" is fully consistent, "MC" is marginally consistent, and "NC" is not consistent. The numbers in each box are the numeric values assigned to each consistency-metric outcome, as detailed in D23B. (b) The consistency score for each experiment over the lead times in (a). (c) The difference between each of the sensitivity experiments and ALL-DROP colored by the relative-impact category scale where "L" is large, "M" is moderate, and "N" is neutral. The table on the bottom right depicts the cutoffs for each category.

[2] The experiments at hand do not allow us to examine R64 in the same manner as other variables for several reasons. First, D23A found that a disparity in observing-system strategy between 2017 and later years likely had strong impacts on R64 results, which complicates analysis. Regardless, these experiments do not address that issue. Further, the relatively smaller sample of R64 reduces the ability to stratify results even at short lead times. Finally, many short lead times for TS have very few cases of R64 during the first few days. Considering the difficulties imposed by these issues, we have elected not to analyze R64 in this paper.

Here, we calculated consistency scores for short- (6–60 h) and long- (66–126 h) lead times, hereafter called "temporal" consistency scores (e.g., Fig. 5b). Then, we took differences in temporal consistency between the experiments and ALL-DROP and assigned each difference to one of five "relative-impact categories" (e.g., Fig. 5c). For the example given in Fig. 5, only sampling the environment (i.e., NO-VOR) would lead to track improvement at short lead times over ALL-DROP. Note that differences between the average MAE skill of the sensitivity experiments and ALL-DROP were also calculated to provide additional context and will be included alongside the differences in temporal consistency in subsequent figures.

A benefit of displaying temporal consistency scores relative to ALL-DROP as in Fig. 5c is that one can immediately identify how consistent results are between experiments. For example, if NO-ENV has degradation relative to ALL-DROP and NO-VOR has improvement relative to ALL-DROP, then it follows that only sampling the environment would benefit the forecast of the variable being analyzed. Conversely, if NO-ENV has improvement relative to ALL-DROP and NO-VOR has degradation relative to ALL-DROP, then only sampling the in-vortex region would benefit the forecast. Other combinations of results are obviously possible, but they require more nuanced interpretation.

Evaluation encompasses the FULL and FULL-W samples at both short and long lead times as well as the three initial classification stratifications for both samples at only short lead times. Initial classification results at long lead times are omitted for three reasons. First, there were generally smaller differences in those results. Second, for outer wind radii in particular, large fluctuations in forecast bias and MAE began at long lead times when the sample size for wind radii decreased to fewer than about 75. In some cases, these fluctuations influenced changes in both MAE skill and consistency in ways that obfuscated interpretation. Finally, the overwater sample size was much smaller than the full sample (around half of the full sample at 120 h), which was too small to stratify at long lead times.

Note that the above framework is most useful when examining NO-IC, NO-ENV, and NO-VOR. Thus, it will be used when discussing results for those experiments in this paper. For ICO, impacts are mostly negative relative to ALL-DROP, and it is more useful to discuss impacts relative to NO-DROP. We therefore discuss the results of ICO in a separate section after detailing results from the other experiments.

3. Results relative to ALL-DROP

This section is divided into discussions of track, intensity, and outer-wind radii forecast results. For each variable, full results are discussed first. Then, the sample is stratified by the initial TC classification (i.e., TS, H12, H345) to illustrate how the impacts of various sampling strategies change through the life cycle of a TC. For more details on these stratifications, see section 2a.

a. Track

Figures 6 and 7 depict the impact of dropsondes on TC track forecasts and serve as a template for many of the remaining figures. Figure 6 depicts the MAE, MAE skill, as well as a consistency scorecard for all experiments. Note that the MAE skill and consistency scorecard are both computed with respect to NO-DROP. Figure 7 then depicts differences in temporal consistency scores as well as differences in time-averaged MAE skill for both the full and overwater samples. As described in section 2c, differences are computed with respect to ALL-DROP to highlight how changes to the observing system impact results.

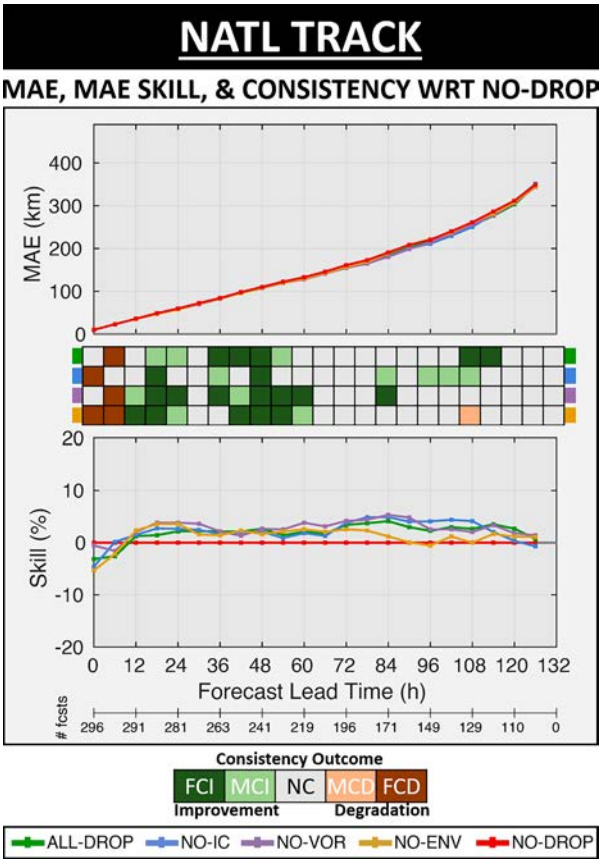

FIG. 6. The MAE, MAE skill, and the consistency metric of the full sample for ALL-DROP (green), NO-IC (blue), NO-VOR (purple), NO-ENV (yellow), and NO-DROP (red). Note that for each experiment, skill was calculated with respect to (WRT) to NO-DROP. Boxes between the MAE and MAE skill panels that use the consistency outcome shading indicate the forecast lead times where results were fully consistent, marginally consistent, or not consistent based on the methodology described in D23B. The sample size is given below the *x* axis.

On average, observing-system changes impacted TC track forecasts mostly at shorter lead times in the full sample (Figs. 6 and 7a). Though dropsondes anywhere improved the track forecast, environmental dropsondes benefited forecasts the most (Fig. 7a). In particular, only sampling the environment (i.e., NO-VOR) improved the temporal consistency of track forecasts relative to ALL-DROP at short lead times in FULL.

Only verifying overwater cases changed the results somewhat, but it did not change the interpretation (Fig. 7c). Specifically, the benefit of sampling only the environment (i.e., NO-VOR)

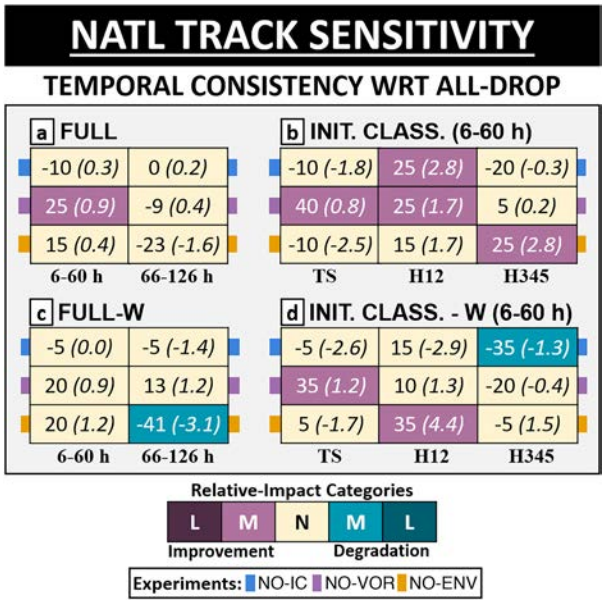

FIG. 7. Relative-impact scorecards for (a),(c) FULL and FULL-W at short and long lead times and (b),(d) FULL and FULL-W stratified by initial classification at short lead times. Included are results for NO-IC (blue), NO-VOR (purple), and NO-ENV (yellow). The sample sizes for FULL (FULL-W) are 296 (283) at 0 h and 219 (156) at 60 h while the sample size for TS, H12, and H345 for the full (overwater) sample are, respectively, as follows: 92, 107, and 94 (87, 103, and 90) at 0 h and 61, 79, 77 (41, 55, and 54) at 60 h. Each cell includes temporal consistency scores calculated with respect to (WRT) to ALL-DROP as well as average MAE skill scores WRT to ALL-DROP in parentheses.

decreased slightly at short lead times to a neutral relative impact. Yet the cost of not sampling the environment (i.e., NO-ENV) increased considerably at long lead times. As such, removing environmental sampling moderately degraded the track forecast for overwater cases. One reason why NO-ENV performed relatively worse in FULL-W than FULL might be because overwater cases are farther from the data-rich United States observation network. That is, it could be that environmental sampling becomes more important for storms in data-sparse regions. Nevertheless, these results again highlight the importance of environmental sampling for TC track forecasts.

Stratifying short-term results by initial classification revealed that the most beneficial regions for dropsonde sampling changed throughout the TC life cycle (Figs. 7b,d). In general, it was best to sample the environment and not the vortex for TS. More specifically, entirely removing in-vortex dropsondes (i.e., NO-VOR) moderately improved TS temporal track forecast consistency relative to ALL-DROP in both the full and overwater samples. Conversely, removing inner-core dropsondes in H345 moderately degraded short-term H345 track forecasts in FULL-W (i.e., NO-IC; Fig. 7d). Note that the suggested improvement due to removing environmental dropsondes in FULL seems to be a result of land impacts, as the signal disappears in FULL-W.

The above results are qualitatively similar to those of several recent studies involving reconnaissance data. For example, Sellwood et al. (2023) found that adding high-resolution inner-core data significantly improved track forecasts of Maria (2017), which was a category-3 hurricane at the time. Meanwhile, Sippel et al. (2022) found that the impacts of adding reconnaissance data into the NCEP GFS model evolved as TCs intensified. In particular, adding flight-level reconnaissance observations (that mostly concentrate over the vortex; their Figs. 7 and 8) improved track forecasts for hurricanes more than they improved forecasts of tropical storms. In one of their examples, they showed that the added data significantly degraded track forecasts of Hurricane Dorian (2019) when it was a TS as it approached the Lesser Antilles, but similar data significantly improved the forecasts a few days later when Dorian was a hurricane and approaching the United States. Section 5 discusses possible reasons why in-vortex dropsondes might degrade track forecasts in weaker TCs.

b. *Intensity*

This section examines the impacts of the sensitivity experiments on TC intensity forecasts. Though one can examine intensity in terms of both VMAX and PMIN, the major focus here will be on VMAX since PMIN-forecast MAE varied much less among the experiments.

1) VMAX

Removing dropsondes anywhere degraded short-term VMAX forecasts in both the full and overwater samples (Figs. 8a and 9a,c). More specifically, VMAX forecast errors at short lead times in all experiments suffered compared to ALL-DROP (Figs. 9a,c). Note that both the relative impact scores and differences in averaged MAE skill degraded more in FULL-W than FULL. This reflects landfall acting to drive the forecast intensity toward a similar weak value in all experiments (i.e., it minimizes the differences). The MAE increased the most when removing all in-vortex dropsondes (i.e., NO-VOR). In particular, NO-VOR demonstrated a large degradation in temporal consistency and a 6.6% and 9.7% decrease in average MAE skill relative to ALL-DROP for FULL and FULL-W, respectively. Much of this degradation was associated with a larger negative VMAX forecast bias (i.e., a weaker TC, Fig. 10a) than the other experiments. Meanwhile, changes to dropsonde sampling had smaller and generally less consistent impacts on VMAX at longer lead times.

In-vortex dropsondes appear to have been the most important for predicting intensity regardless of the initial classification. More specifically, removing in-vortex dropsondes (i.e., NO-VOR) led to the most degradation of temporal consistency, though for major hurricanes removing only inner-core dropsondes had almost as large of an impact. These results are qualitatively similar to a large body of previous research that has shown the importance of thoroughly sampling the TC vortex for VMAX forecasts (e.g., Zhang et al. 2009, 2011; Weng and Zhang 2012; Aberson et al. 2015).

Changes to dropsonde sampling impacted VMAX forecasts of initially weaker TCs the most in both FULL and FULL-W (Figs. 9b,d). In particular, moderate to large degradations of temporal consistency occurred at short lead times if removing any in-vortex dropsonde observations in TS or TS-W. The

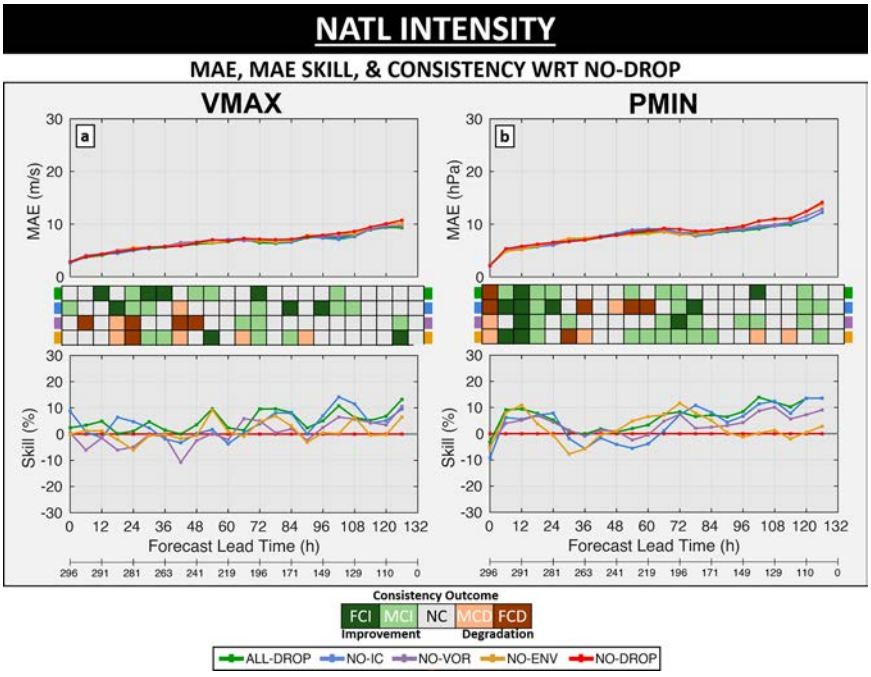

FIG. 8. As in Fig. 6, but for TC forecasts of (a) VMAX and (b) PMIN.

degradation at short lead times was accompanied by a decrease in average MAE skill relative to ALL-DROP in each experiment, ranging from 5.0% to 18.1%. These results highlight the importance of sampling tropical storms, especially within the vortex.

The various experiments affected VMAX forecast biases of initial tropical storms and hurricanes much differently (Figs. 10a–d). Removing dropsondes anywhere in a TS (Fig. 10b) generally made TCs forecasts weaker at short lead times. The greatest impact on the bias came from removing environmental dropsondes. Meanwhile, forecasts initialized from at least hurricane intensity (Figs. 10c,d) clearly produced the weakest storms at short lead times when removing in-vortex dropsondes (i.e., NO-VOR).

2) PMIN

Denying dropsondes had less impact on PMIN than on VMAX, though the impact was larger in FULL-W than FULL (Figs. 8b and 9e,g). For example, only denying inner-core

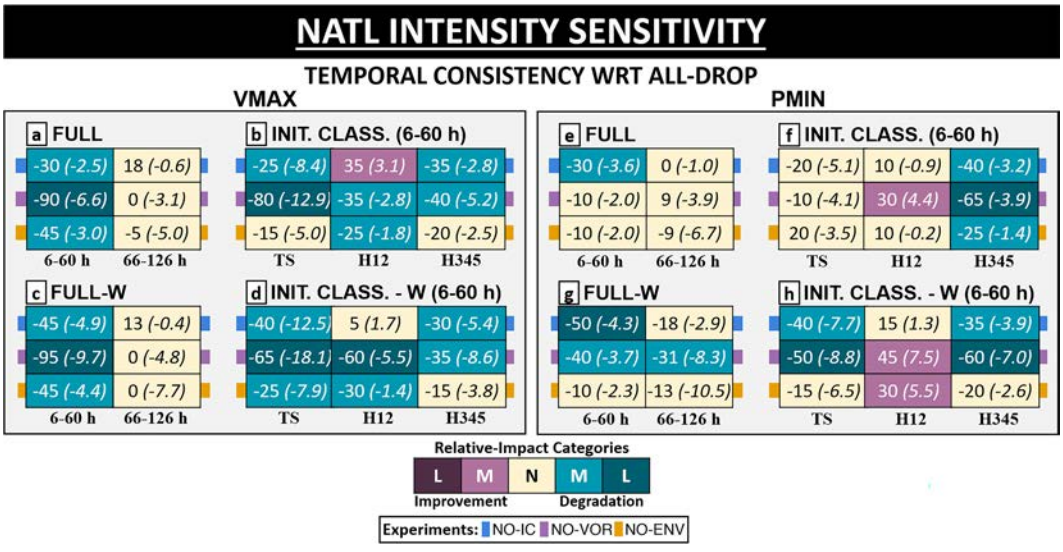

FIG. 9. As in Fig. 7, but for TC forecasts of (a)–(d) VMAX and (e)–(h) PMIN.

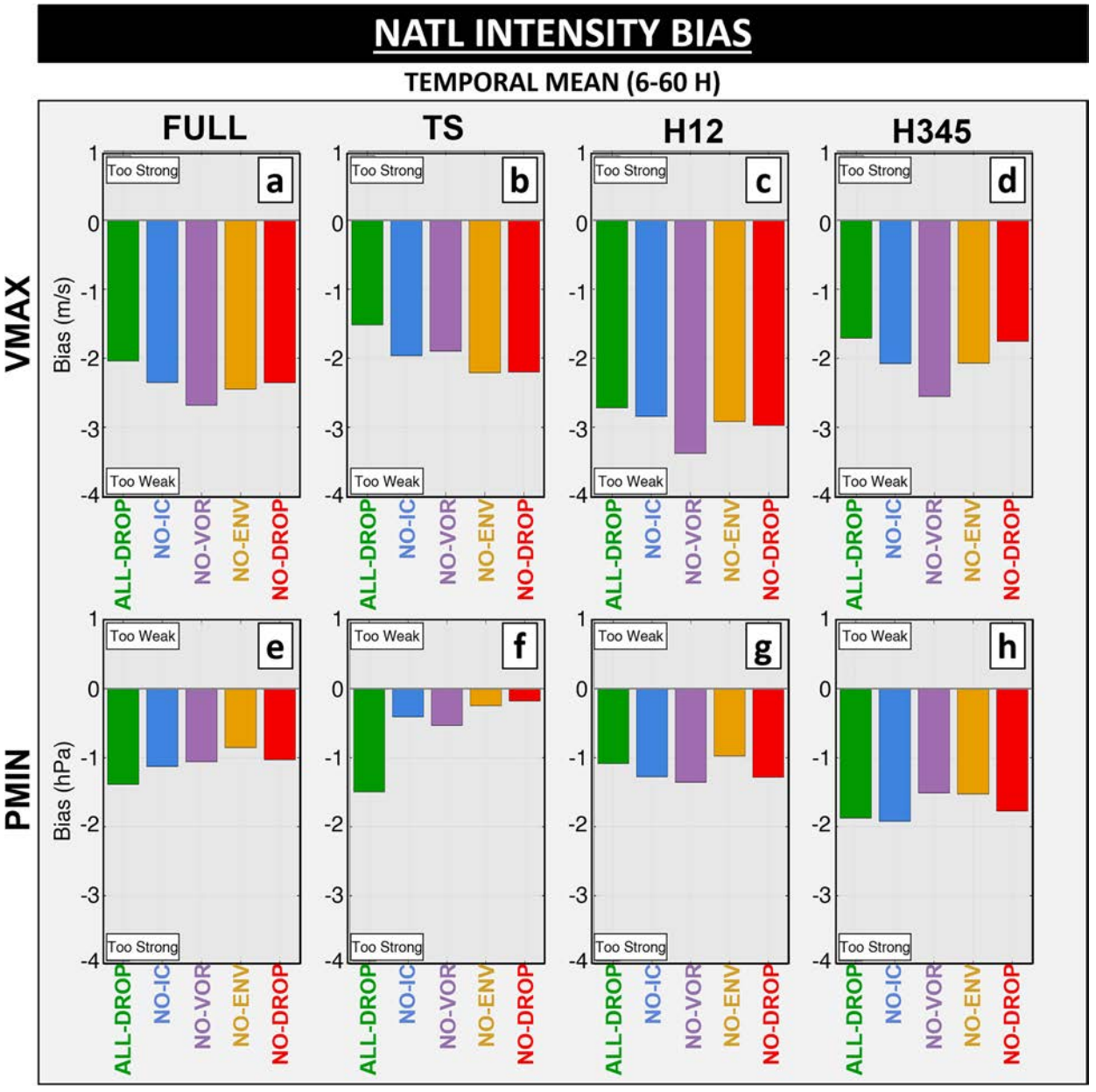

FIG. 10. The mean bias of NATL TC intensity forecasts between 6 and 60 h (i.e., temporal mean bias) for (top) VMAX and (bottom) PMIN both (a),(e) overall and by initial classification into (b),(f) TS; (c),(g) H12; and (d),(h) H345 for ALL-DROP (green), NO-IC (blue), NO-VOR (purple), NO-ENV (yellow), and NO-DROP (red).

dropsondes (i.e., NO-IC) meaningfully changed the temporal consistency in FULL at short lead times (Fig. 9e). For FULL-W, denying dropsondes anywhere over the vortex was detrimental (Fig. 9g). As with VMAX, both the relative impact scores and averaged MAE skill values were more negative for FULL-W compared to FULL.

Stratifying by initial classification reveals that in both the full and overwater samples, H345 drove most of the degradation noted at short lead times (Figs. 9f,h). Removing dropsondes anywhere over the vortex was detrimental to short term H345 in both samples. In the overwater sample, TS also contributed to the degradation seen in NO-VOR.

The differing results for PMIN and VMAX probably reflect an inaccurate pressure-wind relationship in HB20. For example, the VMAX bias in ALL-DROP was negative (i.e., too weak; Fig. 10a), and removing dropsonde data worsened the bias by making the forecast TCs even weaker. Inconsistent with the VMAX bias, the PMIN bias was also negative in ALL-DROP (i.e., too strong; Fig. 10e) so that removing dropsonde data improved the PMIN bias. These contradicting results are

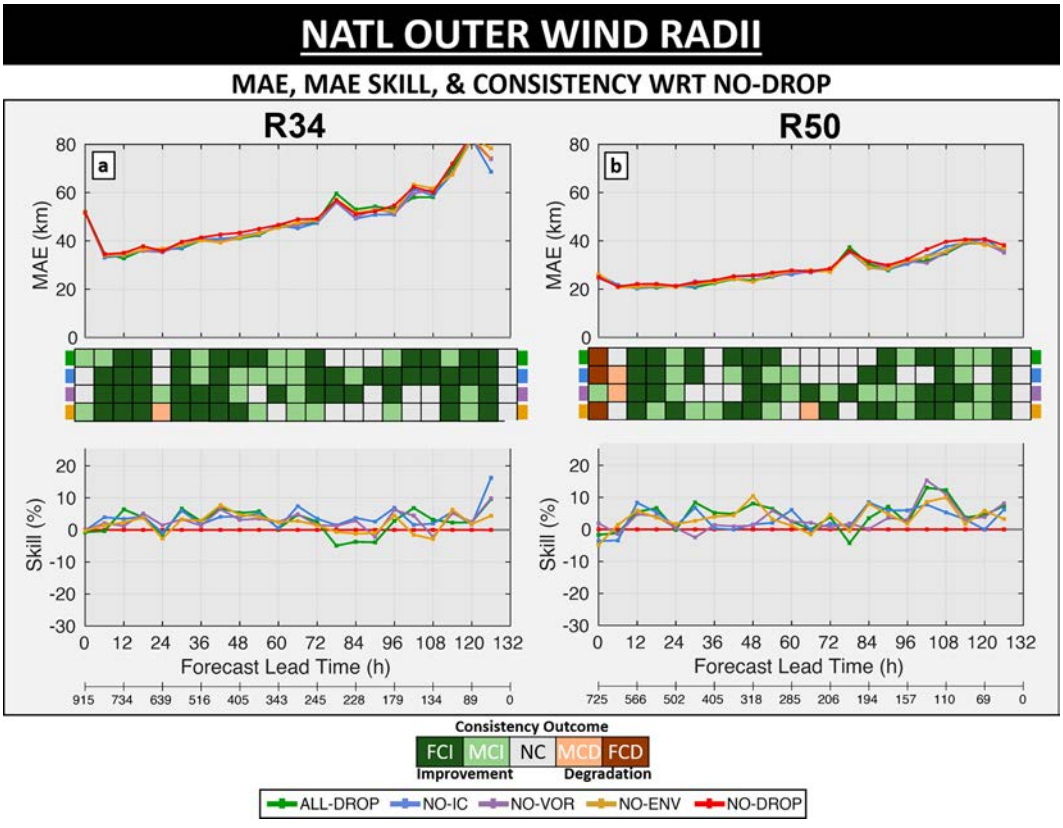

FIG. 11. As in Fig. 6, but for TC forecasts of (a) R34 and (b) R50.

fundamentally similar to those in Lu and Wang (2023), who found that assimilating inner-core radar data improved VMAX and degraded PMIN in an experimental version of HWRF.

c. *Outer wind radii*

This section examines the relative impacts of dropsondes on forecasts of the outer wind radii (R34 and R50). While on average results were similar among experiments, differences emerged when stratifying by initial classification.

1) R34

Dropsondes improved R34 in ALL-DROP with substantial consistency over many lead times, and the sensitivity experiments showed that sampling anywhere similarly benefited short-term R34 forecasts (Figs. 11a and 12a,c) by improving a negative forecast size bias (Fig. 13a). More specifically, short-term R34 forecasts did not appreciably differ from ALL-DROP in either the full or overwater samples so long as dropsondes sampled either the vortex or the environment (Figs. 12a,c). One likely reason for this is that the R34 distribution spanned the boundary between the vortex and the environment (Fig. 4c), so dropsonde observations in either location would sample near R34 in many cases. The degradation driven by inner-core dropsondes at longer lead times in Fig. 12a seems to be a reflection of landfall since it does not occur in the FULL-W sample.

Stratifying by initial classification revealed more nuanced impacts than suggested by the overall sample (Figs. 12b,d). While the sensitivity experiments did not meaningfully change short-term R34 forecast MAE on average, removing dropsondes in most regions degraded short-term TS forecasts in both the full and overwater samples. On the other hand, Figs. 13b–d shows that having dropsondes anywhere improved the negative size bias for hurricanes more so than for tropical storms. This was particularly true for H345 R34 forecasts, where the large negative bias in NO-DROP was reduced between 3 and 3.75 km.

2) R50

The average R50 results behaved similarly to R34 (Figs. 11b and 12e,g). As with R34, dropsondes improved R50 forecasts with substantial consistency across many lead times. ALL-DROP improved upon a negative size bias seen without dropsondes, and the experiments here showed that sampling anywhere with dropsondes likewise improved the bias (Fig. 13a). The improvement in FULL when the vortex was not sampled (i.e., NO-VOR; Fig. 12e) again appears to reflect land interaction, as the signal did not show up in FULL-W.

As with R34, stratifying by initial classification revealed more nuanced results (Figs. 12f,h). In general, as TCs intensified, the location of the most beneficial dropsondes for R50 migrated radially outward from the center in both the full and

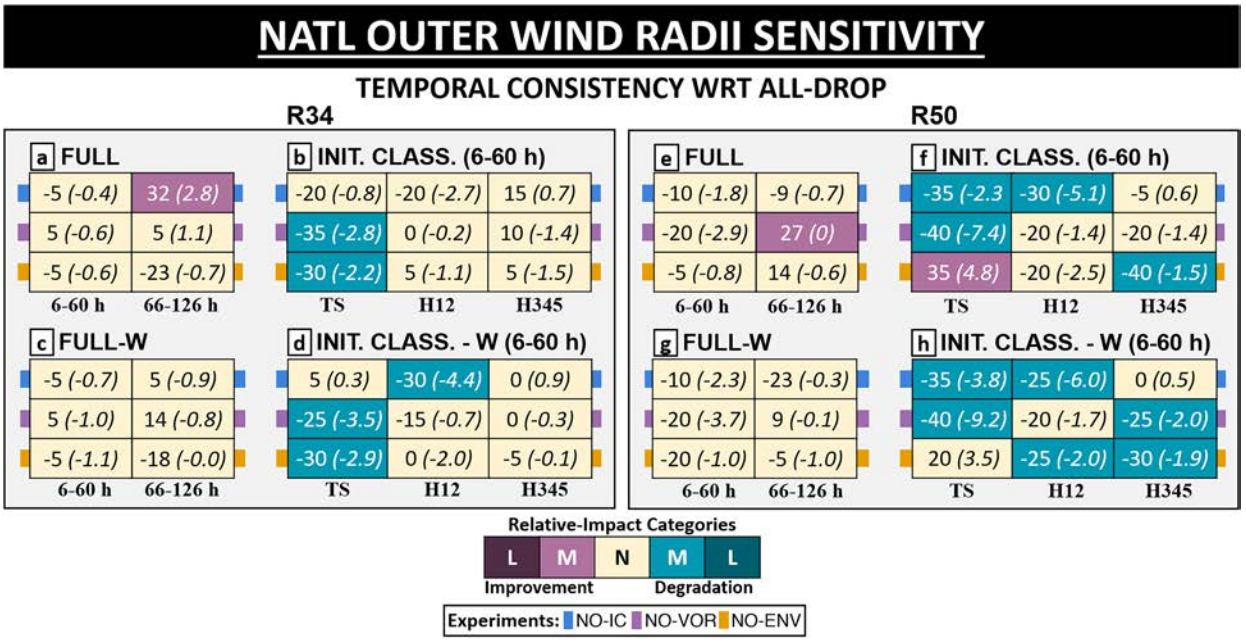

FIG. 12. As in Fig. 7, but for TC forecasts of (a)–(d) R34 and (e)–(h) R50 at short lead times. The sample sizes for R34 for FULL (FULL-W) are 915 (871) at 0 h and 343 (291) and 60 h while the sample size for TS, H12, and H345 for the full (overwater) sample are, respectively, as follows: 210, 353, and 352 (197, 338, and 336) at 0 h and 107, 73, and 160 (87, 63, and 124) at 60 h. The sample sizes for R50 for FULL (FULL-W) are 725 (694) and 0 h and 285 (247) at 60 h while the sample size for TS, H12, and H345 for the full (overwater) sample are, respectively, as follows: 55, 318, and 352 (52, 306, and 336) at 0 h and 81, 48, and 156 (67, 44, and 124) at 60 h.

overwater samples. First, in-vortex dropsondes were most important for TS forecasts of R50. More specifically, removing in-vortex dropsondes (i.e., NO-VOR) in TS moderately degraded the temporal consistency and substantially decreased the MAE skill in both samples relative to ALL-DROP. For H12, the categorical change in consistency bordered the neutral/moderate threshold for both in-vortex and environmental dropsondes in both samples. Comparatively, removing in-vortex dropsondes had less of an impact on H345 (i.e., NO-VOR). Additionally, removing environmental dropsondes (i.e., NO-ENV) degraded the temporal consistency of R50 in H345 more than it did weaker TCs. This signals that environmental dropsondes were more important for stronger TCs, likely since the observed R50 distribution in stronger TCs is at a larger radius than in weaker TCs (cf. Figs. 4f–l). Finally, similar to the results for R34, dropsondes anywhere improved negative R50 forecast biases in hurricanes more than in tropical storms (Figs. 13e–h).

4. Impacts of inner-core-only sampling

ICO relates directly to operationally implemented changes in reconnaissance sampling that occurred after the 2017 hurricane season (see D23A, their Fig. 3). Beginning in 2018, USAF C-130 missions began systematically releasing dropsondes at the end points of their alpha-pattern formations (around 150–200 km from the TC center). Prior to this change, the C-130 dropsondes primarily only sampled the inner 75 km of TCs to estimate PMIN and VMAX. Thus, ICO serves as a rough proxy[3] for the pre-2018 dropsonde strategy employed on USAF flights. Considering that USAF has historically flown much more frequently than NOAA, C-130 missions have often sampled TCs alone, and the ICO sampling configuration for dropsondes frequently occurred in the past. Given the addition of end-point dropsondes on C-130 flights, inner-core-only sampling has rarely taken place since 2018.

ICO results are shown in a different manner than the preceding results. As described in section 2c, ICO impacts are mostly negative relative to ALL-DROP, and it is therefore more useful to discuss the impacts of only sampling the inner core (i.e., ICO) relative to not sampling at all (i.e., NO-DROP). Thus, Fig. 14 depicts scorecard graphics displaying only consistency-metric results overall and stratified by initial classification.

ICO sampling degraded some important aspects of hurricane forecasts. In particular, despite benefiting track forecasts in TS, ICO degraded H12 track forecasts relative to NO-DROP with marginal consistency on days 4–5. Further, intensity forecasts in H345 substantially suffered at most long lead times. In fact, the H345 intensity forecasts in ICO had 5%–10% less skill than NO-DROP after 72 h (not shown). Recall though that sampling the inner core was important for hurricane intensity forecasts when other dropsondes were present (Fig. 9). Thus, these ICO results indicate that having inner-core dropsondes alone degrades intensity forecasts, yet sampling the inner-core along with other

[3] ICO included dropsonde observations not just from the USAF C-130, but also from other reconnaissance aircraft as well (see section 2b).

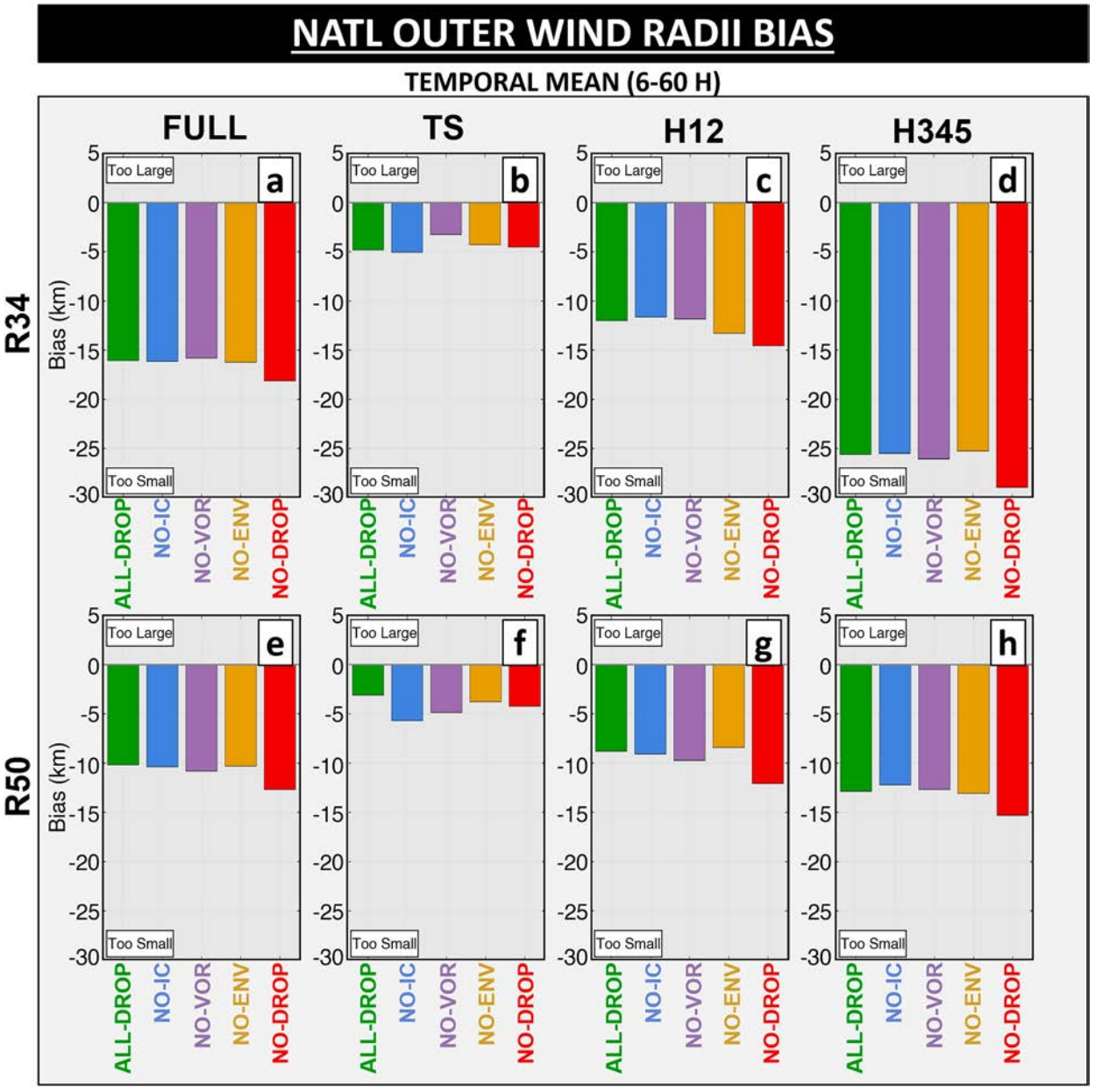

FIG. 13. As in Fig. 9, but for TC forecasts of (a)–(d) R34 and (e)–(h) R50.

regions of the TC benefits intensity forecasts. Further, while the ICO strategy tended to degrade forecasts of outer wind radii in TS, it actually improved those forecasts in hurricanes, particularly H345. Nevertheless, considering the track or intensity degradation across many lead times for hurricanes, ICO results suggest that only sampling the inner core with dropsondes is undesirable. Further, these results suggest that adding end-point dropsondes in 2018 likely benefited forecasts for hurricanes.

5. Summary and conclusions

This work represents the most comprehensive assessment to date of the relative impact of TC dropsonde observations in the TC inner core, in-vortex region, and environment on TC forecasts. It follows the framework of Ditchek et al. (2023a) (D23A), which assessed the overall impacts of dropsondes in the 2017–20 NATL hurricane seasons following standard NHC forecast verification procedures (Cangialosi 2022). More specifically, it uses the 2020 version of the basin-scale, multistorm configuration of HWRF (HB20) to conduct four observing-system sensitivity experiments that deny dropsonde data in various annuli. Except for which dropsondes were denied, the experimental setup here is identical to D23A. Thus, the experiments can be compared to the D23A ALL (here called ALL-DROP) and NO (here called NO-DROP) experiments, effectively creating a group of six experiments.

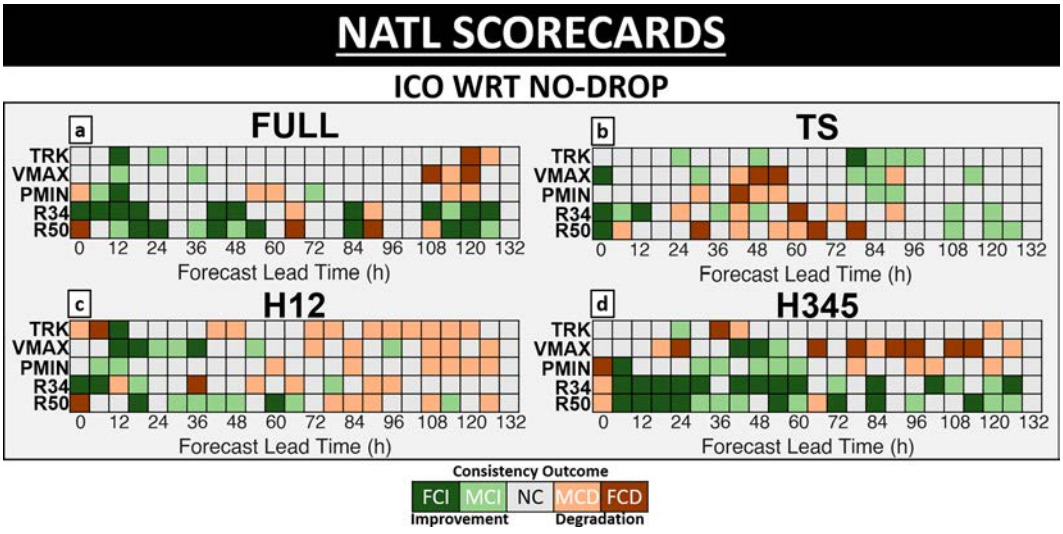

FIG. 14. Consistency scorecards for NATL TC forecasts (a) overall and stratified by initial classification into (b) TS, (c) H12, and (d) H345. The sample sizes for track (TRK), VMAX, and PMIN for FULL, TS, H12, and H345 are, respectively, as follows: 296, 92, 107, and 94 at 0 h; 219, 61, 79, 77 at 60 h; and 110, 30, 27, and 50 at 120 h. The sample sizes for R34 for FULL, TS, H12, and H345 are, respectively, as follows: 915, 210, 353, and 352 at 0 h; 343, 107, 73, and 160 at 60 h; and 89, 16, 20, and 50 at 120 h. The sample sizes for R50 for FULL, TS, H12, and H345 the sample sizes are, respectively, as follows: 725, 55, 318, and 352 at 0 h; 285, 81, 48, and 156 at 60 h; and 69, 12, 10, and 47 at 120 h.

Analysis focused on the subsample of D23A individual forecasts with direct dropsonde sampling that used HWRF-cycled mesoscale error covariance during DA (i.e., OBS-HCOV from D23A). This choice was guided by D23A, which concluded that sampling TCs with dropsondes can directly improve TC forecasts only if using sufficiently advanced DA techniques (cf. their Fig. 21c that included forecasts that used HWRF-cycled mesoscale error covariance to their Fig. 21d that included forecasts that used global-model error covariance). This study then stratified results by initial TC classification to explore whether and how the impact of different dropsonde-sampling strategies changed throughout the TC life cycle. Results were shown for both a full sample and an overwater sample, where forecasts of TCs that were over land or lead times where TCs were forecasted to be overland were removed.

Because this study compares results from six different experiments, we have taken a few steps to compare results in a tractable manner. As a starting point for subsequent analysis, this

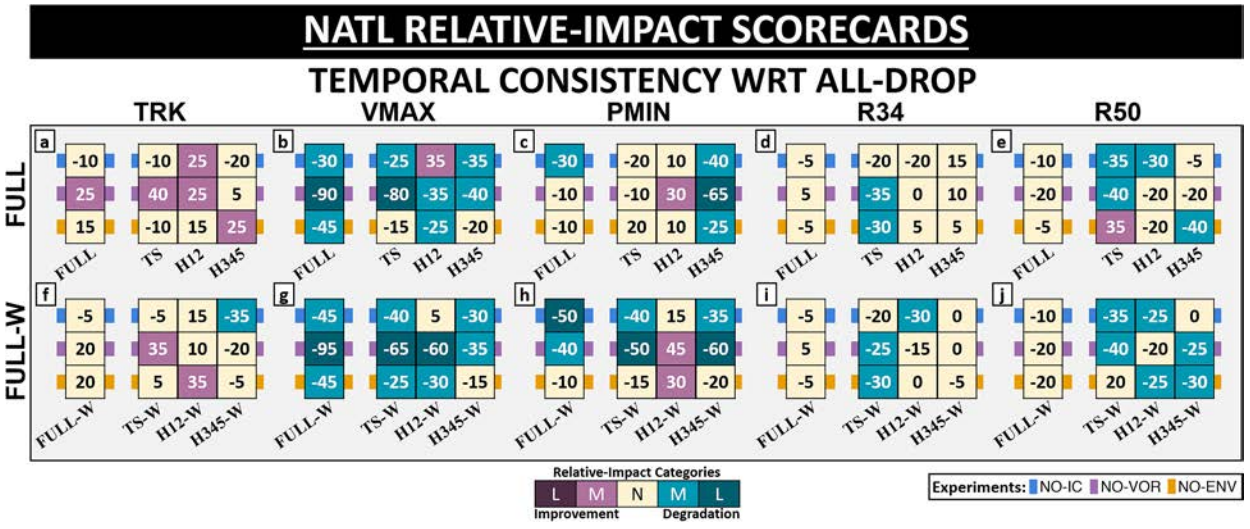

FIG. 15. Summary graphic of the relative impact of the temporal consistency of each experiment compared to ALL-DROP for both the full and overwater samples. Note that these results were previously displayed and are here organized by variable for NO-IC (blue), NO-VOR (purple), and NO-ENV (yellow) at short lead times (i.e., 6–60 h).

paper uses a new metric developed by Ditchek et al. (2023b) (D23B) and used in D23A. This metric—the consistency metric—helps identify when improvement or degradation occurs consistently and not because of a handful of outliers, which is a common problem in TC verification (see D23B for more details). The consistency-metric results are then used to evaluate the performance of each experiment at short and long lead times relative to ALL-DROP. While the analysis is admittedly high-level, the approach significantly improves the clarity of the results.

Before summarizing results, it is useful to briefly discuss the main takeaways from D23A that are relevant to this study (i.e., comparing ALL and NO from their OBS-HCOV sample). Overall, D23A found that dropsondes directly improved forecasts of track, intensity, and outer wind radii (their Fig. 21c). The impacts on outer wind radii were particularly positive and extended through most lead times. Similar results are found in this study when comparing ALL-DROPS to a NO-DROPS baseline (Figs. 6, 8, and 11), albeit for a smaller sample than D23A (see section 2c).

To aid in summarizing the main results found in this study, Fig. 15 depicts relative-impact scorecards for all forecast variables assessed. The figure summarizes results for three of the four experiments (i.e., NO-IC, NO-VOR, and NO-ENV). Note that ICO is not included as forecasts generally were degraded relative to ALL-DROP (i.e., forecasts were degraded when dropsonde observations outside of the inner core were denied). Additionally, since the most interesting and useful results occurred at short lead times, the graphic excludes long lead times.

An immediately obvious result in Fig. 15 is that removing dropsondes anywhere substantially degraded VMAX forecasts (Figs. 15b,g). The worst forecast performance occurred when removing in-vortex sampling (i.e., NO-VOR), especially for TS. Further, considering that the largest number of dropsonde data in H345 falls outside 250 km (Fig. 4k), these results suggest that bringing sampling of major hurricanes inward somewhat could improve forecasts of their intensity. These results agree with the large body of research that has previously shown the benefits of assimilating inner core data (e.g., Zhang et al. 2009, 2011; Weng and Zhang 2012; Aberson et al. 2015). Nonetheless, the degradation found when only sampling the in-vortex region (i.e., NO-ENV) suggests that sampling the environment also benefited VMAX forecasts. This suggests that synoptic surveillance missions conducted by the NOAA G-IV, which are the primary source of environmental dropsondes, benefit VMAX forecasts. A likely reason for this benefit is that the deep-layer G-IV dropsondes sample otherwise unobserved aspects of the near-TC environment. For example, though satellite-based retrievals such as atmospheric motion vectors yield valuable data, they typically give little information regarding the wind field below about 400 hPa near TCs [e.g., Fig. 6 of Lim et al. (2019) and Fig. 3 of Li et al. (2020)]. Without such observations, analyses can miss important details such as the profile of vertical wind shear.

In-vortex dropsondes provide several intensity-dependent benefits in addition to substantially improving VMAX forecasts. For example, sampling the in-vortex region improved forecasts of TS outer wind radii (Figs. 15d,e,i,j). Though the importance of in-vortex sampling decreased considerably for R34 forecasts in hurricanes, it generally remained important for R50 forecasts in all TCs. This difference was likely due to an outward-shifting distribution of the outer wind radii as TCs intensified. While R34 in hurricanes spanned the boundary between the vortex and the environment (cf. Figs. 4f–i,l), R50 always fell within the vortex. An additional benefit of in-vortex sampling was its substantial impact on PMIN in H345 (Figs. 15c,h). Given the relationship of PMIN to TC structure and damage (e.g., Chavas et al. 2017; Klotzbach et al. 2020), this suggests that dropsondes within the in-vortex region of a major hurricane are particularly beneficial.

Figure 15 also demonstrates that track forecasts in weaker TCs benefited more from environmental sampling, while forecasts in stronger TCs benefited more from in-vortex sampling (Figs. 15a,f). Though changing the sampling strategy from environment-focused to vortex-focused as a TC intensifies could benefit track forecasts, this would increase forecast MAE for both intensity and outer wind radii in weaker TCs (Figs. 15b–e,g–j). Alternatively, degraded track forecasts when in-vortex data are added to weaker TCs could suggest a deficiency in DA. Tropical storms are known to be more asymmetric than hurricanes, and it is certainly plausible that asymmetric analysis increments in weaker systems could degrade the track by projecting onto larger scales than they should. While such investigation is beyond the scope of this study, improving DA methods to constrain increments to the appropriate scales could improve the impact of dropsondes, particularly those in the vortex, on the track forecasts. One possible example of such a method has been provided in Huang et al. (2021), who showed that scale-dependent localization in the GSI framework can improve track forecasts.

Another result from this study is that only sampling the inner-core region should be avoided. More specifically, ICO routinely performed worse than ALL-DROP (not shown) and even degraded forecasts relative to NO-DROP (Fig. 14). Long lead time degradation relative to NO-DROP occurred for H12 track, intensity, and outer-wind-radii forecasts as well as H345 intensity forecasts. These results have a great deal of operational relevance because USAF flights did not release dropsondes outside the inner core prior to 2018. Thus, the ICO configuration roughly represents the pre-2018 operational dropsonde sampling strategy when NOAA missions were not present. In this situation, it would have been better to not assimilate dropsonde data at all rather than to only assimilate the inner core dropsonde data available from the USAF. Fortunately, that strategy has been used infrequently since USAF missions began routinely releasing dropsondes at the ends of their radial legs (about 150–200 km from the TC center) in 2018. Results here suggest that change has likely benefited TC forecasts.

Future work should explore how dropsondes can optimally be used with other airborne data, particularly within the TC vortex, to improve TC forecasts. The results here suggest that substantially increasing sampling with dropsondes in certain regions could further benefit forecasts, though the ideal number of dropsondes may not be operationally practical. For example, forecasts of outer wind radii improve when they are sampled, echoing the finding from D23A that sampling the near-core

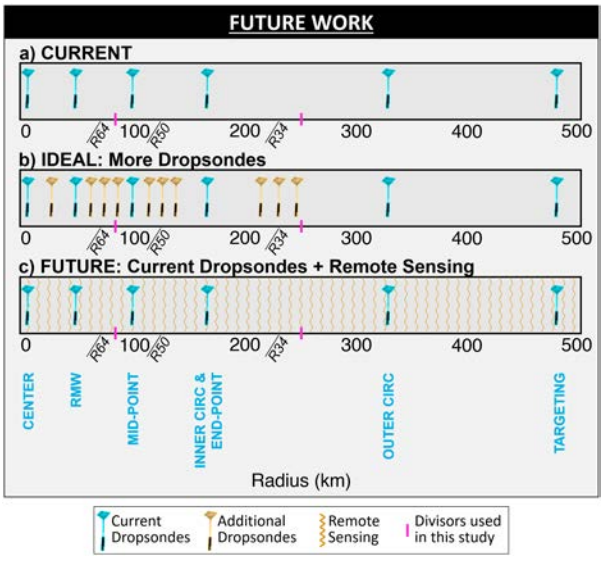

FIG. 16. Summary graphic of the (a) current, (b) ideal, and (c) more realistic future dropsonde-sampling strategy, with current dropsondes in blue, additional dropsondes in gold, remote sensing in gold, and divisors used in this study in pink. Also included is the mean location of R34 and R50 as found in Fig. 4c and the corresponding mean location of R64.

region is important for predicting the hurricane-force wind field. While adaptive dropsonde sampling that targets outer wind radii could further improve forecasts, a potential problem with this approach is that it vastly increases the number of dropsondes deployed (e.g., Fig. 16b). This is problematic not only given the substantial cost associated with acquiring more dropsondes (around $800 each; M. Brennan 2022, personal communication), but also since launching that many dropsondes from a single aircraft mission is physically and logistically challenging with current technology. A superior long-term approach might be to assimilate remotely sensed reconnaissance data that continuously samples the wind and thermodynamic fields (e.g., Fig. 16c) in addition to the current density of dropsonde data (i.e., Fig. 16a). Indeed, in recent years the Imaging Wind and Rain Airborne Profiler (IWRAP; Guimond et al. 2014) and the Doppler wind lidar (DWL; Bucci et al. 2018; Bucci 2020; Zhang et al. 2018) have routinely been deployed on the NOAA P3s and have been used to retrieve accurate, high-resolution profiles within (IWRAP) and outside of (DWL) precipitation. Further research needs to quantify how this additional data can be used with currently assimilated data to improve TC forecasts. Thermodynamic data are more challenging, particularly within precipitation, though recent experiments with airborne radio occultation have shown promise (Haase et al. 2021).

Finally, though this study has yielded valuable comparisons of the relative impacts of inner-core, in-vortex, and environmental dropsondes, more work remains to improve the current dropsonde-sampling strategy. Two ongoing studies are assessing sampling strategies for G-IV synoptic surveillance missions. One study is examining the impacts of dropsondes launched during the inner circumnavigation, a flight pattern that was added in 2018 (see D23A, their Fig. 3c). The second study is assessing the impact of all G-IV reconnaissance, reconnaissance from the G-IV inner circumnavigation, and reconnaissance from the G-IV environmental targeting. Results found from D23A, this study, and these future 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 ID NA19OAR0220188). Thank you to AOML/HRD's Xuejin Zhang and Sundararaman (Gopal) Gopalakrishnan for allocating additional resources for the experiments, Andrew Hazelton, Gus Alaka, and three WAF reviewers (Chris Landsea and two anonymous) for their constructive comments on the manuscript, and EMC and DTC (especially Evan Kalina, Zhan Zhang, Linlin Pan, Biju Thomas, and Mrinal Biswas) as well as Gus Alaka 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 ID 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.

——, 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.

——, A. Aksoy, K. J. Sellwood, T. Vukicevic, and X. Zhang, 2015: Assimilation of high-resolution tropical cyclone observations with an ensemble Kalman filter using HEDAS: Evaluation of 2008–11 HWRF forecasts. *Mon. Wea. Rev.*, **143**, 511–523, https://doi.org/10.1175/MWR-D-14-00138.1.

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

Alaka, G. J., Jr., 2019: 2019 basin-scale HWRF (HWRF-B): An HFIP real-time demonstration project on WCOSS. *HRD Monthly Science Meeting*, Miami, FL, NOAA, https://noaahrd.wordpress.com/2019/11/19/hrd-monthly-science-meeting-of-november-2019/.

——, 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.

——, 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.

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

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

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

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

Cangialosi, J. P., 2022: National Hurricane Center forecast verification report: 2021 hurricane season. NHC Tech. Rep., 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.

Ditchek, S. D., J. A. Sippel, G. J. Alaka Jr., S. B. Goldenberg, and L. Cucurull, 2023a: A systematic assessment of the overall dropsonde impact during the 2017–20 hurricane seasons using the basin-scale HWRF. *Wea. Forecasting*, **38**, 789–816, https://doi.org/10.1175/WAF-D-22-0102.1.

——, ——, P. J. Marinescu, and G. J. Alaka Jr., 2023b: 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.

Goldenberg, S. B., S. G. Gopalakrishnan, V. Tallapragada, T. Quirino, F. Marks Jr., 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.

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

Haase, J. S., M. J. Murphy, B. Cao, F. M. Ralph, M. Zheng, and L. Delle Monache, 2021: Multi-GNSS airborne radio occultation observations as a complement to dropsondes in atmospheric river reconnaissance. *J. Geophys. Res. Atmos.*, **126**, e2021JD034865, https://doi.org/10.1029/2021JD034865.

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 III, 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.

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.

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

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

Lu, X., and X. Wang, 2023: A study of simultaneous assimilation of coastal ground-based and airborne radar observations on the prediction of Harvey (2017) with the hourly 3DEnVar system for HWRF. *J. Geophys. Res. Atmos.*, **128**, e2022JD037681, https://doi.org/10.1029/2022JD037681.

——, ——, Y. Li, M. Tong, and X. Ma, 2017: 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.

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.

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., J. A. Sippel, and A. Aksoy, 2023: Assimilation of Coyote small uncrewed aircraft system observations in Hurricane Maria (2017) using operational HWRF. *Wea. Forecasting*, **38**, 901–919, https://doi.org/10.1175/WAF-D-22-0214.1.

Simpson, R. H., and H. Saffir, 1974: The hurricane disaster—Potential scale. *Weatherwise*, **27**, 169–186, https://doi.org/10.1080/00431672.1974.9931702.

Sippel, J. A., 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. D., 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.

Velden, C. S., and S. B. 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., 90–93.

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

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

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.

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

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

Zhang, 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.