Evaluation of Hurricane Analysis and Forecast System (HAFS) Error Statistics Stratified by Internal Structure and Environmental Metrics

GEORGE R. ALVEY, III,[a,b] GHASSAN J. ALAKA JR.,[b] LEW GRAMER,[a,b] AND ANDREW HAZELTON[a,b]

[a] *University of Miami, CIMAS, Miami, Florida*
[b] *NOAA/OAR/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida*

(Manuscript received 20 February 2024, in final form 14 September 2024, accepted 23 October 2024)

ABSTRACT: This study quantifies tropical cyclone (TC) error statistics from the Hurricane Analysis and Forecast System (HAFS) across different environmental conditions (e.g., vertical wind shear) and inner-core structural metrics. A particular focus is the evolution of poorly understood aspects of internal TC structure, including vortex tilt, and their impact on forecast errors. Although previous studies have demonstrated that vortex tilt, vertical wind shear, and precipitation processes impact TC intensity and track, this is the first known study to stratify these cooperative interactions to gain insights into their relationships with forecast errors. A 3-yr retrospective sample of forecasts in the North Atlantic basin from two HAFS configurations (HAFS-A and HAFS-B) demonstrates that TCs with larger tilt magnitudes have larger forecast track errors on average than smaller tilt TCs. Smaller tilt magnitudes have larger absolute intensity errors in short-range forecasts, whereas larger tilt magnitudes tend to have larger negative intensity biases at medium range. TCs with a tilted vortex are shown to have both left-of-shear [maximizing downshear left (DSL)] and left-of-tilt-oriented positional track biases. Furthermore, those cases with greater downshear biases tend to have more convection and larger positive intensity biases, highlighting the importance of the interplay between inner-core characteristics and forecast errors.

KEYWORDS: Hurricanes/typhoons; Tropical cyclones; Forecast verification/skill; Forecasting techniques; Numerical weather prediction/forecasting

1. Motivation/introduction

The confidence in a particular model forecast can vary depending on the characteristics of environmental conditions surrounding tropical cyclones (TCs). Zhang and Tao (2013) and Tao and Zhang (2015) demonstrated in idealized simulations that predictability decreases with increasing values of vertical wind shear (VWS) but only to a certain point when high VWS causes development to cease altogether. Bhatia and Nolan (2013) quantified the effects of synoptic variables and environmental conditions on track and intensity forecast errors from the National Hurricane Center (NHC), dynamical models, and statistical models using a dataset of TCs from 2006 to 2010. They found that although greater values of VWS were associated with larger forecast errors, this effect tended to maximize in moderate VWS regimes (5–12 m s^{-1}). More recent case studies have also partly attributed moderate VWS to decreased predictability (Munsell et al. 2013; Rios-Berrios et al. 2016a,b; Alvey et al. 2020). The large spectrum of horizontal (e.g., radius of maximum winds) and vertical (e.g., vertical tilt or misalignment and vortex height; DesRosiers et al. 2023) vortex structures including precipitation distributions in cases with moderate shear can also decrease the predictability of future intensity change (Tao and Zhang 2015; Alvey et al. 2020). Furthermore, Sumwalt et al. (2017) have identified moderate VWS cases as "some of the most significant challenges for tropical cyclone forecasting and are prone to causing large forecast errors."

Trabing and Bell (2020) expanded upon Bhatia and Nolan (2013) by examining NHC forecast error stratifications of environmental parameters from the Statistical Hurricane Intensity Prediction Scheme (SHIPS) in both the Atlantic and east Pacific basins using a larger and updated dataset from 1989 to 2018. They found, on average, broader error distributions for TCs that undergo rapid intensification (RI) within climatologically favorable environments. Although the larger error distributions they found in favorable environments differ from the largest errors in moderate VWS found by other studies like Bhatia and Nolan (2013), their focus on RI (and rapid weakening) cases (rather than all cases) and differing samples can likely explain some of these discrepancies. Perhaps more importantly, Trabing and Bell (2020) hypothesized that internal storm dynamics, which remain poorly understood, are likely the cause of their finding that the largest RI errors occur in favorable environments.

This study addresses those poorly understood aspects of internal storm dynamics by uniquely examining how both environmental conditions and internal storm characteristics relate to forecast errors. The importance of achieving an aligned vortex for RI to occur has been well documented within the literature (Frank and Ritchie 2001; Zhang and Tao 2013; Rios-Berrios et al. 2016b, 2018; Chen et al. 2019; Alvey et al. 2020, 2022; Alvey and Hazelton 2022; Stone et al. 2023). Misaligned storms are commonly weaker TCs initially (minimal hurricane strength or weaker) and may be in moderate VWS environments. However, determining whether misaligned storms will align, and the timing of such events, remains challenging (Finocchio and Majumdar 2017; Munsell et al. 2017; Yu et al. 2023), hence one of the difficulties in improving RI forecasts.

Corresponding author: George R. Alvey, george.alvey@noaa.gov

DOI: 10.1175/WAF-D-24-0030.1

Despite these challenges, some recent studies have established potential relationships between alignment and RI with TC size. For example, Alvey et al. (2020) found that normalizing the tilt by the radius of maximum winds (RMWs) improved the correlations of tilt as a predictor with future intensity change, and a ratio of tilt/RMW < 0.75 was necessary for rapid intensification in an ensemble of simulations from Edouard (2014) in 28°–30°C SST. Schecter (2022) further tested this result in idealized experiments with varying SSTs wherein he found ratios of 0.31 ± 0.12 for 26°C, 0.54 ± 0.21 for 28°C, and 0.74 ± 0.24 for 30°C SSTs preceded by sustained alignment and more substantial intensification. These relationships, however, and their impacts on forecast errors remain poorly understood. Thus, this study places an emphasized focus on vortex tilt and its relationships with precipitation and intensity change in the context of associated forecast errors. The goal is to achieve a better understanding of forecast errors to help model developers prioritize potential problem areas for future upgrades and to help forecasters understand when they should or should not place confidence in a particular model solution.

2. Data and methods

a. HAFS model configurations

This study uses the Hurricane Analysis and Forecast System (HAFS) operational versions "A" and "B" with storm-centric nested configurations (a 6-km static nest and a 2-km storm-following nest; Hazelton et al. 2023). The primary differences between HAFS-A and HAFS-B exist within the physical suites, wherein HAFS-A uses the Geophysical Fluid Dynamics Laboratory (GFDL) microphysical parameterization (Chen and Lin 2013; Zhou et al. 2022) and HAFS-B uses the Thompson microphysical scheme (Thompson et al. 2008). HAFS-B also uses the "tc-pbl" TC-specific adjustments to the eddy-diffusivity mass-flux (EDMF)–TKE planetary boundary layer (PBL) scheme (Chen et al. 2022, 2023). Retrospective simulations covering the Atlantic basin from 2020 to 2022 with identical configurations (including initial and boundary conditions) are run by NOAA's Atlantic Oceanographic and Meteorological Laboratory (AOML)/Hurricane Research Division (HRD) in collaboration with the University of Miami and Cooperative Institute for Marine and Atmospheric Studies (CIMAS) and Environmental Modeling Center (EMC).

The verification of HAFS uses the NHC best track for information on TC intensity in terms of both the maximum-sustained 10-m wind and the minimum central pressure, as well as TC position estimates every 6 h over the duration of the TC lifetime (Landsea and Franklin 2013).

All TCs in HAFS are tracked using the GFDL vortex tracker (Marchok 2021). For the purposes of this study, all 3-hourly forecast output periods during, and within 6 h of, the TC center traversing any landmass[1] are removed from the sample. All subsequent forecast outputs for a given cycle are also removed following land interaction, as defined above. Forecast periods with a verification position overland are also removed. Additionally, only those simulations after which the storm has been designated as a TC[2] (including tropical depressions) are included in the dataset.

b. Environmental and internal storm dynamic metrics

Variables from the retrospective simulations are calculated for each 3-hourly output time following conventions from the SHIPS predictor files (DeMaria et al. 2005). This study primarily utilizes deep-layer VWS magnitude (SHRD, 850–200 hPa within 200–800 km of the storm center) and direction (SHTD), midlevel VWS (SHRS and SHTS, 850–500-hPa VWS magnitude and direction within 200–800 km of the storm center), and 200–800-km midlevel (700–500 hPa, RHMD) relative humidity.

In addition to environmental variables, internal storm dynamic characteristics like vortex tilt are also calculated for each storm's output time. TC centers are calculated for each vertical level in 25-hPa increments following methods from Nguyen et al. (2014) by using geopotential height centroids. Vortex tilt is defined as the maximum displacement between the low-level center at a height of 800 hPa (~2.0 km) and any TC center in the 525–425-hPa (~5–6.5-km) layer, hereafter referred to as 2–6.5-km tilt. Unless otherwise stated, the tilt magnitudes for all initial conditions use the 6-h forecasts to account for large initial adjustments (0–3 h) caused by potential inconsistencies between vortex initialization and inner-core data assimilation processes. The RMW is calculated as the distance between the maximum azimuthally averaged wind at 2 km and the GFDL tracker surface center estimate, which assumes that there is no material difference between wind fields and corresponding centers in the lower troposphere. Although the axisymmetric RMW may introduce a low bias in highly asymmetric TCs, this methodology was chosen due to the noisiness in alternative point-based methodologies (Hazelton et al. 2023). The TC precipitation structure is also partitioned into either convective modes or stratiform modes based on a modified algorithm[3] using 2-km model reflectivity (Steiner et al. 1995; Rogers 2010).

c. Overview

Figure 1 demonstrates the distributions of the 6-h forecasts for 1171 simulations from HAFS-B and 1225 from HAFS-A that have no land interaction. The intensity histogram (Fig. 1a) shows that the majority of TCs at 6 h after initialization time are tropical storm (TS) strength with a median intensity of 51 kt (1 kt ≈ 0.51 m s^{-1}) in HAFS-B and 50 kt in HAFS-A. The VWS in Fig. 1b using the SHIPS SHRD spans from 10.3 to 23.7 kt (10.4–23.0 kt) within the interquartile range (IQR) of HAFS-B

[1] The landmass filter includes all major landmasses and island chains.

[2] TCs designated as subtropical or extratropical are not included.

[3] A more detailed description of the partitioning algorithm tuning for HAFS can be found in Hazelton et al. (2021).

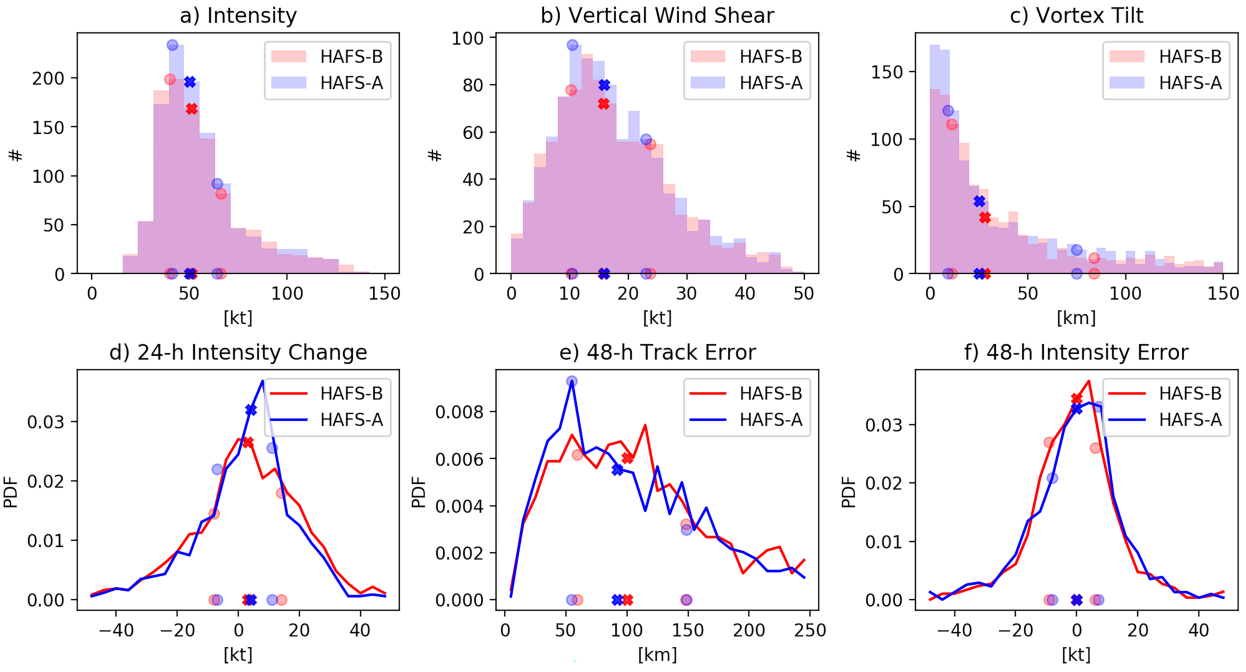

FIG. 1. Histogram distributions (number of samples) for HAFS-B (red) and HAFS-A (blue) of (a) 6-h intensity, (b) VWS (SHRD, 200–800 km 850–200 hPa), and (c) vortex tilt (2–6.5 km). Probability density function (PDF) for (d) 24-h intensity change (kt), (e) 48-h track error (km), and (f) 48-h intensity error (kt) for HAFS-B (red) and HAFS-A (blue). The "X" symbols correspond to the median, and the circles denote the 25th and 75th percentile with the color matching each respective model. Each PDF has 25 bins which equates to a bin increment of 4 kt in (d), 10 km in (e), and 4 kt in (f).

(HAFS-A). Nearly half of HAFS-B (HAFS-A) simulations initialize with a vortex tilt (2–6.5 km, Fig. 1c) greater than 27.9 km (25 km) using methodologies described in section 2b. HAFS-A simulations tend to have slightly lower tilt magnitudes with a 25-km median and smaller tilt magnitudes within the IQR. The 24-h intensity change shown in Fig. 1d generally follows a normal distribution; however, the IQR is slightly skewed toward greater intensity change rates in HAFS-B (−8 to +14 kt) than in HAFS-A (−7 to +11 kt). These positive skews are potentially related to the removal of land interaction cases, which are mostly weakening TCs; those systems that undergo extratropical transition or dissipate (via rapid weakening) are also not included in the verification. The 48-h track error distribution (Fig. 1e) has a 113-km mean and 100-km median error and IQR from 59 to 148 km in HAFS-B, which also closely corresponds to the NHC mean errors from 2017 to 2021 (114 km; Cangialosi 2022). The HAFS-A 48-h track errors are comparatively small with a 110-km mean, a 92-km median, and an IQR from 54 to 149 km. The HAFS-B 48-h intensity bias (Fig. 1f) IQR from −9 to +6 kt indicates a slight skew toward underprediction of intensification, which can also likely be attributed to missed RI events. On the other hand, the HAFS-A 48-h intensity error IQR ranges from −8 to +7 kt indicating little or no bias. Although the mean absolute 48-h intensity errors of 10.5 kt in HAFS-B and 10.8 kt in HAFS-A improve upon the 2017–21 NHC average error of 10.9 kt, it is important to consider that NHC forecasts are also commonly based on the early runs[4] [e.g., HAFS-A early cycle (HFAI) and HAFS-B early cycle (HFBI)] or previous cycles of numerical models. Furthermore, the cited NHC forecast errors account for all cases including land interaction, whereas the HAFS forecasts examined in this study do not include land interaction cases. Analyses hereafter that only show one model configuration will focus on HAFS-A, wherein HAFS-B has qualitatively similar results.

3. Results

a. *Relationships between tilt and intensity (change) in HAFS*

This section shows relationships between internal storm characteristics (e.g., vortex tilt) and external environmental factors (like VWS) and the overall intensity (change) of TCs. The HAFS retrospective dataset presents a unique opportunity to explore these relationships within a large sample of real case simulations for the first time and provide comparisons to a recent investigation of the relationships with observational datasets like the tail Doppler radar (TDR; Fischer et al. 2023b).

In addition to the removal of cases with land interaction and sub- or extratropical designation as described in section 2a, the analysis in Fig. 2 removes high-latitude TCs (>35°N) that

[4] "Early runs" use the previous model run (e.g., 0600 UTC) and adjust the forecast based on the current position and intensity (e.g., 1200 UTC) of a TC.

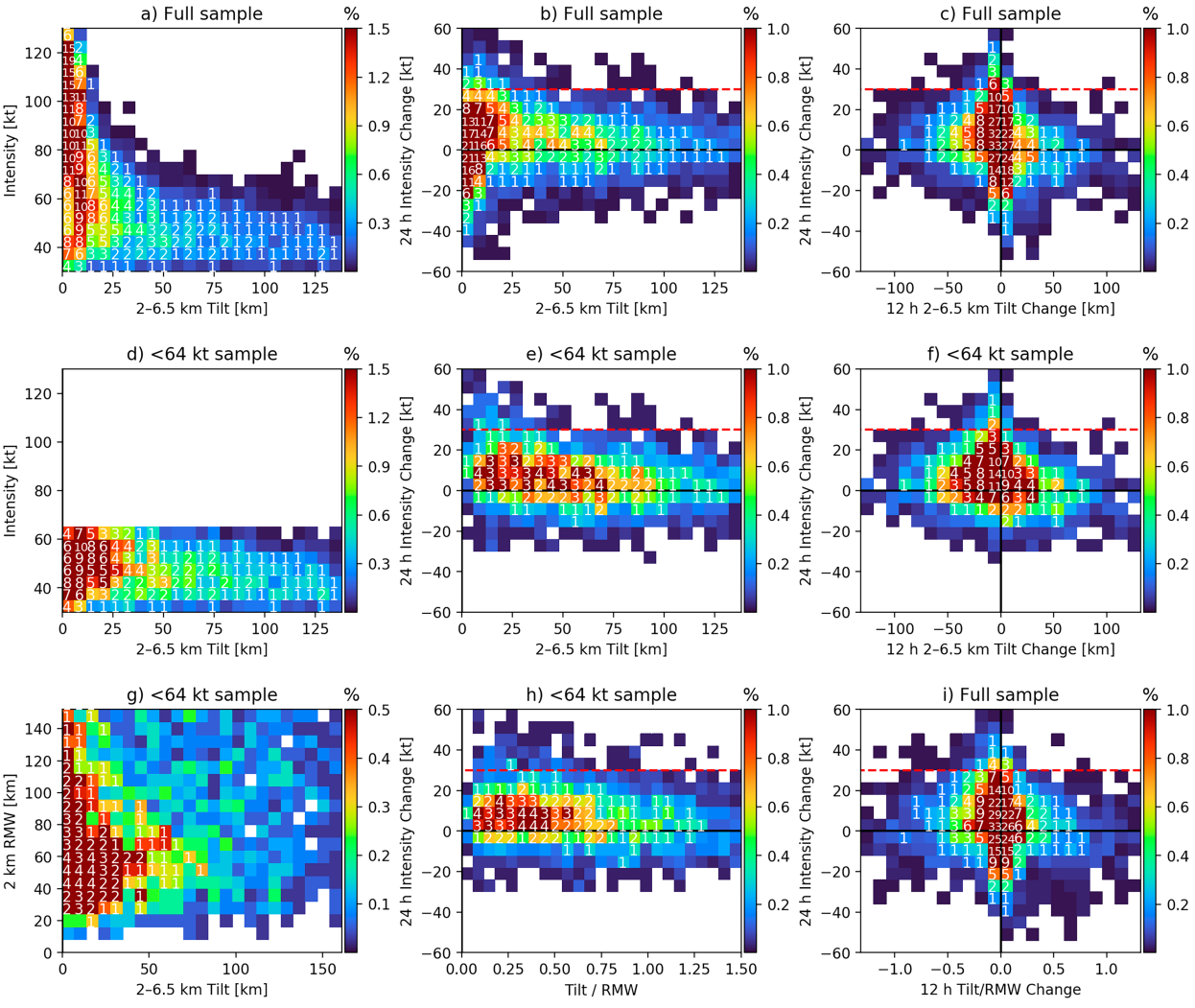

FIG. 2. (a) The relationship between HAFS-A 2–6.5-km tilt (km) and intensity (kt), (b) tilt and future 24-h intensity change (kt), (c) and future 12-h tilt change and future 24-h intensity change for all forecast time periods at a latitude less than 35°N shown in percent distribution (color fill) with sample sizes (white text, ×10). Bins with less than 10 samples do not have numbers. (d)–(f) The configurations follow the top three panels except they are restricted to only those cases < 64 kt. (g) The relationship between tilt and 2-km RMW (km) and (h) the ratio of tilt divided by 2-km RMW and future 24-h intensity change for only those cases < 64 kt. (i) The future 12-h change in the ratio of tilt/RMW and future 24-h intensity change for all cases. All panels with a 24-h intensity change [*y* axis; panels (b), (c), (e), (f), (h), and (i)] utilize a time-averaged vortex tilt (as defined in section 2b but averaged from −6 to +6 h for a given forecast hour using 3-h outputs). The red dashed lines in (b), (c), (e), (f), (h), and (i) delineate the RI threshold (30 kt+/24 h).

are often recurving, undergoing extratropical transition, or traversing lower-ocean heat content. Stronger TC intensity (>64 kt) has a robust negative correlation with vortex tilt (Fig. 2a), wherein a small vortex tilt < 30 km is necessary for intensity with maximum sustained winds greater than 85 kt. TC intensity < 64 kt, however, has a much weaker relationship with vortex tilt (Fig. 2d), a result also shown by Fischer et al. (2023b). Those TCs less than hurricane strength range from 0 to 150+ km vortex tilt magnitudes. Therefore, it is important to understand the internal dynamics responsible for differing vortex tilts in these TCs, how they become virtually aligned (<~25 km; Alvey et al. 2020, 2022) to attain strong hurricane status, and any potential relationships with future intensity change.

Because several previous case studies and modeling studies have highlighted the importance of vortex alignment for future TC intensity change (Frank and Ritchie 2001; Reasor and Eastin 2012; Rappin and Nolan 2012; Tao and Zhang 2014; Nguyen and Molinari 2015; Rios-Berrios et al. 2016a, 2018; Chen et al. 2019; Rogers et al. 2020; Rios-Berrios 2020; Alvey et al. 2020; Schecter and Menelaou 2020; Alvey et al. 2022; Alvey and Hazelton 2022; Fischer et al. 2023a,b; Nam et al. 2023), Fig. 2b shows the histogram distribution for vortex tilt and future 24-h intensity change. Only those panels with 24-h intensity change (*y* axis, Figs. 2b,c,e,f,h,i) utilize an average vortex tilt parameter (tilt as defined in section 2b but averaged from −6 to +6 h for a given forecast hour between

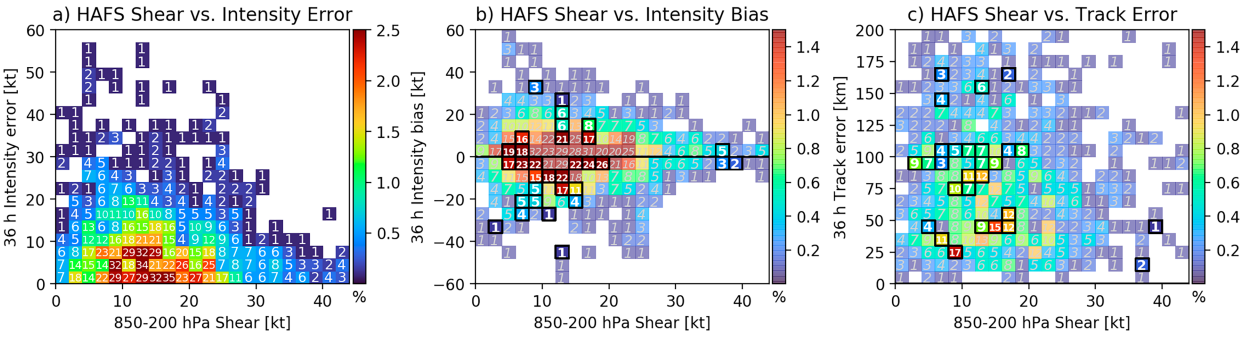

FIG. 3. As in Fig. 2, but showing the relationship between (a) VWS at 6 h (850–200 hPa 200–800-km average, SHRD) and 36-h intensity errors, (b) intensity bias, and (c) track error, for all HAFS-A and HAFS-B cases shown in percent distribution (color fill) with sample sizes (white text, ×10). The black boxes with bolded sample size numbers in (b) and (c) indicate all the bins with RI cases.

6 and 114 h using 3-h outputs) to account for high temporal variability. Little or no statistical relationship exists between average vortex tilt and future TC intensity change for the full sample. However, nearly all the cases that undergo RI (defined as a 30 kt or greater intensification in 24 h; Kaplan and DeMaria 2003; red dashed line) have an initial tilt < 40 km, which indicates the importance of a nearly aligned vortex for RI, a result also found by several of the aforementioned studies documenting alignment. Because there are a few outlier cases that undergo RI despite a large vortex tilt (>40 km), an aligned vortex is not always a necessary condition for RI. This can likely be at least partly attributed to cases that have rapidly evolving vortex-scale evolution (e.g., vortex reformation events; Nguyen and Molinari 2015; Chen et al. 2019; Alvey et al. 2022), wherein 3-hourly output may not be sufficient. Figure 2e shows the same intensity change versus tilt distribution but for only those cases that are TS (34–63 kt) or tropical depression (TD; <34 kt) strength. A more apparent relationship between tilt and future TC intensity change emerges in which those TCs with a small tilt tend to have greater intensification rates; however, a large distribution of intensity change outcomes still exists for all vortex tilt magnitudes.

In addition to the relationship between vortex tilt and future TC intensity change, other previous studies have demonstrated that TCs often transition from a misaligned configuration to vertical alignment prior to (Munsell et al. 2017; Leighton et al. 2018; Alvey et al. 2020; Rios-Berrios et al. 2018) or during the early stages of rapid intensification (Chen and Gopalakrishnan 2015; Alvey and Hazelton 2022; Rios-Berrios et al. 2016b). Figure 2c shows the relationship between 12-h tilt changes (0 to +12 h) and future 24-h intensity changes. Overall, a distribution emerges wherein those TCs with larger increases in tilt magnitude tend to weaken more (defined by maximum sustained winds), and storms with larger tilt magnitude reductions tend to intensify. Despite this pattern, though, many storms also do not follow this general trend, since tilt change is also a function of the initial tilt (i.e., TCs with an already small tilt can only reduce their tilt by that amount). If the sample in Fig. 2c is restricted to only those cases with an initial tilt above the upper quartile (75th percentile, 77 km, not shown), then a pattern emerges with many rapidly intensifying cases having tilt reductions (−80 to −30 km). While there are some cases that still intensify despite tilt increases, the large majority undergo a decrease in intensity. Figure 2f shows a similar relationship isolating only those TCs at TS or TD strength but features a slightly weaker correlation (−0.18 Spearman rank correlation coefficient r_s as opposed to −0.24 for the full sample) between tilt change and future TC intensity change.

Figure 2g (bottom left) shows a histogram of 2-km RMW and 2–6.5-km tilt for all TCs less than hurricane strength; although a positive correlation between tilt and RMW (greater tilt correlates with greater RMW) is apparent, a large distribution of tilts exists for a given RMW. The relationship between 24-h intensity change and the ratio of tilt/RMW (Fig. 2h, −0.23 r_s) shows a weaker relationship than 24-h intensity change and 2–6.5-km tilt (as in Fig. 2e, −0.29 r_s). This indicates limited potential value at least for this particular size metric when considering tilt relationships. Figure 2i also has a weaker correlation between tilt/RMW (−0.15 r_s) and 24-h intensity change than just using tilt magnitude (−0.24, Fig. 2c). Although slightly less robust than the relationships in Fig. 2i, previous 12-h tilt/RMW changes do offer predictive value for future 24-h intensity change (−0.15 r_s, not shown).

b. *Forecast errors stratified by environmental parameters*

Despite previous quantifications of the relationships between forecast errors (track and intensity) and environmental conditions for other modeling systems (Bhatia and Nolan 2013) and NHC forecasts (Trabing and Bell 2020), Fig. 3 is the first such demonstration for the HAFS-A and HAFS-B model configurations. HAFS-B has similar distributions to HAFS-A, and no material differences in the distributions exist when examining difference plots between the two modeling configurations for all results shown in Fig. 3. Figure 3a shows the histogram distribution of 36-h intensity error and 850–200-hPa VWS magnitude. Similar to Bhatia and Nolan (2013), large intensity errors (15–25 kt) tend to occur within the moderate-shear regime (8–20 kt; Rios-Berrios and Torn 2017). It is notable, however, that >40% of the largest intensity forecast errors (>25 kt) are within the favorable low-shear regime (<8 kt), which also corroborates Trabing and Bell (2020) who demonstrated that the largest NHC intensity

errors tend to occur within favorable environments (largely due to poor prediction of storms that undergo RI).

Figure 3b shows the same distribution but for 36-h intensity bias to help determine whether the large absolute intensity errors are a result of overprediction or underprediction of intensity by HAFS. Although there are some cases with large negative intensity biases (<−20 kt 36 h), they follow no specific pattern with respect to the VWS other than the fact that all cases with an intensity bias <−20 kt have a VWS magnitude less than 25 kt. Those cases with high VWS (>25 kt) tend to have a higher predictability with small intensity errors. Cases with the largest positive intensity biases (>30 kt at 36 h) tend to be located within favorable VWS environments (<20 kt). Most RI cases (Fig. 3b, black boxes) fall within a low–moderate-shear environment (6–20 kt) with negative intensity biases. The relationships between VWS and intensity bias indicate that those storms with high VWS (>25 kt) may have more predictable short-range intensity forecasts (36 h or less) as the environmental forcing dominates any internal storm dynamical feedbacks that can be more difficult to predict in lower- or moderate-shear environments (wherein alignment and RI may occur). On the other hand, the relationship between 36-h track error and VWS in Fig. 3c shows a much weaker correlation. Interestingly, however, a majority of the largest 36-h track errors (>125 km) fall within favorable (<8 kt) or moderate shear (8–20 kt) regimes.

To further demonstrate the effects that VWS can have on forecast errors, Fig. 4 shows violin plots of the error distributions for HAFS-A (left violins) and HAFS-B (right violins) stratified by forecast hour (x axis) in high-shear (highest 75th percentile, blue shadings) and low-shear (lowest 25th percentile, red shadings) environments. The cases that initialize in high VWS in both HAFS-A (>22.7 kt, dark blue shading) and HAFS-B (>23.4 kt, light blue shading) have larger track errors than those in lower VWS (<10.5 kt, dark red shading for HAFS-A; 10.3 kt, light red shading for HAFS-B) on average only for forecast periods 12–24 and 96–120 h (Fig. 4a). Track error differences between the high-shear and low-shear subsets are not statistically significant[5] at any time period in both HAFS-A and HAFS-B. It is important to note that sample sizes are smaller (32–103 samples per stratification) for medium-range forecast hours (96–120 h), wherein the magnitudes of median and mean track errors greatly diverge between high-shear and low-shear cases for both model configurations. On average, HAFS-B forecasts also tend to have larger track errors than HAFS-A at medium ranges for both the high- and low-shear quartiles; however, the track error differences between HAFS-A and HAFS-B are only significant at 96 h (red plus sign) for the low-shear stratification. Although not included in Fig. 4, the distributions of moderate VWS cases generally fall between the high-shear and low-shear distributions. The poor track predictability of higher VWS cases at medium ranges suggests that the interplay between shear and other internal storm dynamic parameters may also play a critical role in these cases, particularly given the effects that VWS can have on precipitation distributions and vortex structure (like vertical tilt).

Figure 4b shows the intensity error (bias) distributions for high-shear and low-shear subsamples. Although relatively little mean bias is observed in the high-shear and low-shear samples for both models in the first 72 forecast hours, the means of low-shear cases in HAFS-A do have slight negative intensity biases compared to slight positive biases for high-shear cases during the first 24 h, also evidenced by the violin plot shapes (outward protrusions and thickness). HAFS-B, on the other hand, does not follow this pattern and has much smaller mean differences between the high- and low-shear subsamples; however, the differences at 12–24 h are statistically significant. Despite the similar 48–72-h means in both shear subsamples for HAFS-A (left violins) and HAFS-B (right violins), the low-shear cases (red shading) appear to have much larger variability in intensity errors (particularly negative intensity biases from missed RI cases) evidenced by the increased outward protrusion of the violin plot shapes at intensity errors farther away from the mean. At the 72–120 forecast hours, a signal emerges in which both HAFS configurations tend to over intensify TCs in low-shear environments. The high-shear cases tend to have slightly more negative intensity biases for 72–120 h, though these differences are not statistically significant. The larger amounts of both positive and negative bias outliers for the low-shear cases in HAFS-A and HAFS-B indicate a potential lower predictability of intensity (likely attributed to the potential for RI in more favorable conditions).

The amount of midtropospheric dry air surrounding the TC core can affect the impacts of VWS on TC intensity (Tang and Emanuel 2010, 2012). Dry midtropospheric air can be unfavorably transported into a TC inner core by VWS through downdraft ventilation or radial transport (Alland et al. 2021a,b; Fischer et al. 2023b). This can detrimentally weaken convection and/or cause lower-entropy air to "flush" into the boundary layer. Therefore, given the linkages between VWS and forecast errors shown in Figs. 3 and 4, Fig. 5 also relates the midtropospheric (700–500 hPa) relative humidity (RH) to the deep-layer VWS and 36-h intensity biases. Cases with low VWS (<10 kt) tend to have comparatively high midlevel RH (>60%) on average, and cases with high VWS (>20 kt) tend to have lower midlevel RH (<60%) on average. Nearly all cases with large negative intensity biases (with periods of RI) have favorable RH, a result similar to Trabing and Bell (2020) who found that the majority of cases with large, negative intensity biases tend to have favorable environmental conditions prior to undergoing RI. TCs in 10–20-kt (moderate) VWS with large positive intensity biases also have high midlevel RH (>65%) on average. A small subset of storms with 20–25-kt shear and large negative intensity biases (<−30 kt) have somewhat drier midtropospheric RH (<55%), wherein intensification occurred despite forecasted weakening in HAFS.

This section demonstrates that TCs within generally favorable environments characterized by low-to-moderate VWS (<15–20 kt) and midtropospheric humidity > 50%–60% tend

[5] The differences between distributions (of track errors for the high shear and low shear subsets in this example) are statistically significant (denoted by an asterisk) if a Mann–Whitney U test yielded a p value less than 0.05.

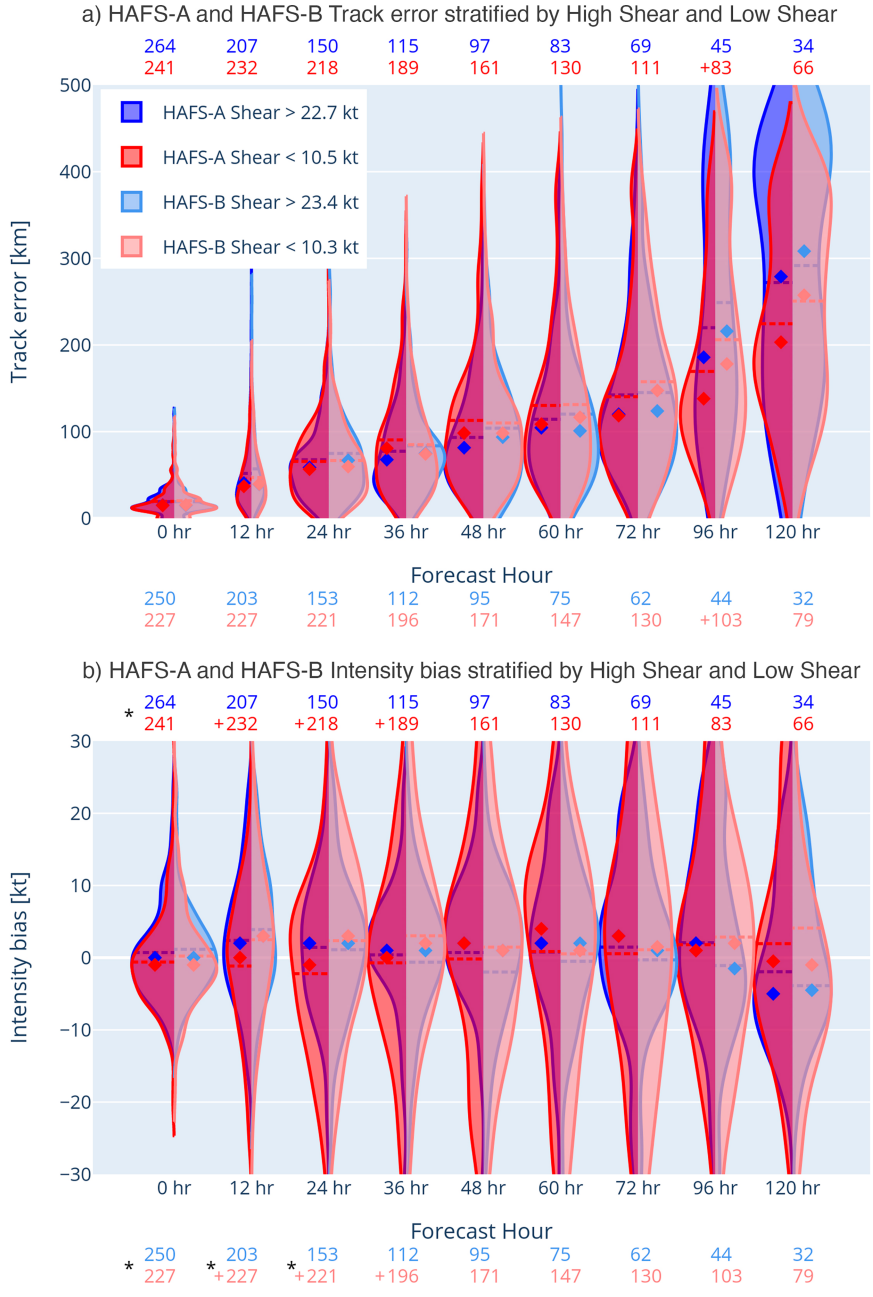

FIG. 4. Full sample violin plots of (a) forecast track error and (b) intensity bias by forecast hours (x axis) stratified by the lowest 25th percentile of 6-h VWS magnitude (red shadings; <10.3 kt, HAFS-B; <10.5 kt HAFS-A) and highest 25th percentile (blue shadings; >22.7 kt, HAFS-A; >23.4 kt, HAFS-B) for HAFS-A (darker shading, left violin) and HAFS-B (lighter shading, right violin). The width of the violins is determined by the approximate frequency of data at a given location with a greater thickness corresponding to the greater frequency. Dashed lines in the middle of violin plots indicate the mean, and diamonds indicate the median. Top numbers and bottom numbers correspond to sample sizes with colors matching the given model and threshold. Asterisks correspond to statistically significant differences (Mann–Whitney U test, 95% confidence interval) between the errors of the lowest and highest 25th percentiles of VWS for HAFS-A (top samples) and HAFS-B (bottom samples). Plus symbols correspond to significant differences between HAFS-A and HAFS-B errors for low VWS (red) and high VWS (blue).

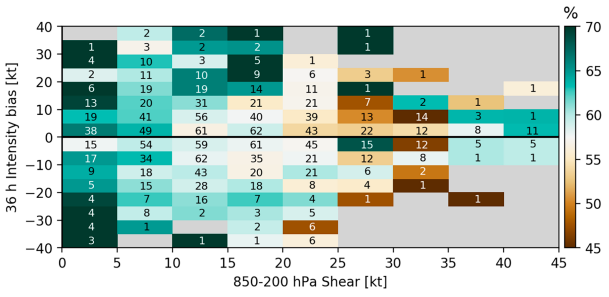

FIG. 5. The relationship between VWS at 6 h (850–200 hPa 200–800 km average, SHRD) and 36-h intensity biases (*y* axis) as a function of the corresponding color-shaded midlevel RH (700–500 hPa 200–800-km average, RHMD) averaged in 5-kt bins for all HAFS-A and HAFS-B cases. Sample sizes for each bin are denoted by the numbers.

to have larger intensity errors than systems in drier or higher VWS environments. The ambiguities in Fig. 5 also provide motivation for section 3d, wherein environmental parameter-based forecast error stratification will be expanded upon by introducing internal storm dynamic parameters like vortex tilt and convective coverage.

c. *Forecast errors stratified by internal storm dynamical parameters*

No previous study to our knowledge has examined the systematic relationships between TC internal and TC dynamical parameters like vortex structure and forecast errors. Trabing and Bell (2020) hypothesized that the large forecast errors associated with TCs that undergo RI in favorable environments are likely the result of internal storm dynamics; however, they were unable to examine this hypothesis in more detail within their study due to limited observations of such variables. The HAFS retrospective simulations examined in this study present a unique opportunity to not only test hypotheses focusing on internal processes from recent observational and modeling studies (as in Fig. 2) but also quantify the potential impacts on the forecast errors. These include relationships with variables and metrics identified as potentially important indicators for future intensity change in other recent studies like vortex tilt (Munsell et al. 2017; Rios-Berrios et al. 2018; Alvey et al. 2020; Alvey and Hazelton 2022; Fischer et al. 2023b), precipitation coverage and symmetry (Jiang 2012; Tao and Jiang 2015; Alvey et al. 2015; Zawislak et al. 2016), and convective intensity (Nguyen and Molinari 2015; Guimond et al. 2016; Alvey et al. 2020; Stone et al. 2023).

The relationship between vortex tilt and forecast track error is demonstrated with violin plots for HAFS-A and HAFS-B in Fig. 6a. The tilt magnitudes for all 6-h forecasts (used as initial condition) without land interaction are delineated into a large tilt (above the 75th percentile) and small tilt (below the 25th percentile) category for both HAFS-A and HAFS-B. These values are 62.6 km (64.9 km) and 8.6 km (9.6 km) for HAFS-A and HAFS-B, respectively (full distributions are shown in Fig. 1c). For both HAFS-A (left violins) and HAFS-B (right violins), the cases with initially large tilt (blue shading) have larger track errors than small tilt cases (red shading) for all time periods. These track error differences are statistically significant for all times in HAFS-A and in HAFS-B except for 96–120 h. Some of these track error differences can be attributed to the larger initial position errors in large tilt cases seen at 0 h. However, a subsample of storms with small initial track errors was stratified into large tilt and small tilt (not shown) and found similar results in the short (6–72 h) and medium (72–120 h) ranges. Furthermore, the track error differences in Fig. 6a may also be related to differences in the TC intensity. TCs in the large tilt subset have a mean initial intensity of 44 kt in HAFS-A (42 kt, HAFS-B), and TCs in the small tilt subset have a 72-kt (77 kt, HAFS-B) mean initial intensity, which also reflects the distribution shown in Fig. 2a. The shapes of the violin plots for forecast hours 24–36 h indicate that the track error differences between large tilt (larger track errors in large tilt cases) and small tilt cases are not primarily caused by outliers, also evidenced by the smaller displacement between means and medians than seen at later forecast periods. Interestingly, the gap between median forecast errors of large tilt and small tilt is larger for HAFS-A than for HAFS-B at most forecast times. Storms with larger tilt magnitudes also have slightly greater mean errors (dashed lines) in HAFS-A than in HAFS-B through 36 h; however, at 120 h, HAFS-B has larger mean errors with both large and small tilt samples. It is important to note that due to the smaller sample sizes at 96–120 h (~200 total cases or less), the impacts on these medium-range time periods should be tested further in future studies with a larger sample size.

Because large mean initial intensity differences exist between the tilt subsets in Fig. 6a and the largest variability of tilt occurs in weaker TCs (Fig. 2a), Fig. 6b stratifies to only include those TCs with maximum sustained winds less than 64 kt. The 0-h track errors for this subsample, overall, are larger than the full sample in Fig. 6a, though the differences between large tilt and small tilt for HAFS-A and HAFS-B are now negligible. The results are similar to Fig. 6a, wherein large tilt cases have larger track errors on average than small tilt cases for most time periods; however, these differences are much smaller than the full sample and are only statistically significant at the 95% confidence interval at 12 h (in HAFS-A).

In addition to the relationship of vortex tilt with forecast track errors (as demonstrated by Fig. 6), intensity errors are shown in Fig. 7. Figure 7a shows a similarly configured violin plot for absolute intensity errors (kt) separated into the same large tilt and small tilt stratifications as Fig. 6a. For both modeling configurations, small tilt cases have larger average absolute intensity errors than large tilt cases (0–36 h) that are statistically significant at 0–12 h in HAFS-A (0 h in HAFS-B); the opposite pattern was seen from the stratifications with track forecast errors (Fig. 6). This is likely at least partly attributed to most RI periods occurring in small tilt cases (Figs. 2b,e). It is interesting to note, however, that at medium ranges (72–120 h) the average absolute error becomes larger (statistically significant at 72–120 h in HAFS-A and 96–120 h in HAFS-B) for cases with large initial tilt in both modeling configurations. One possible explanation is cases that transition from large tilt to small tilt and rapidly intensify. This is more readily apparent in the intensity bias (Fig. 7b), wherein

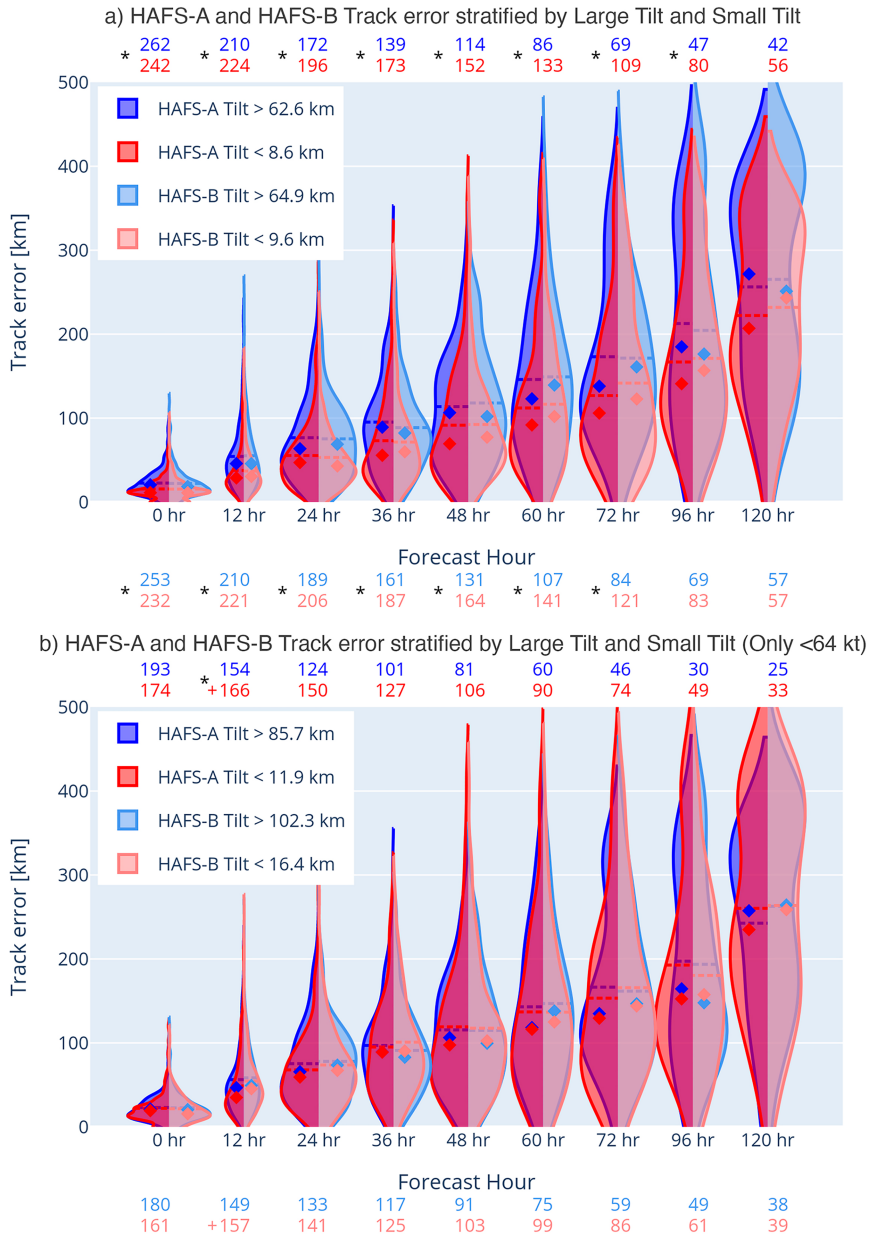

FIG. 6. As in Fig. 4, but for (a) full sample and (b) only storms less than 64-kt violin plots of track error (km) by forecast hours stratified by the lowest 25th percentile of 6-h 2–6.5-km tilt (red shadings; <8.6 km, full sample HAFS-A; <11.9 km, 64-kt sample HAFS-A; <9.6 km, full sample HAFS-B; <16.4 km, 64-kt sample HAFS-B) and highest 25th percentile (blue shadings; >62.6 km, full sample HAFS-A; >85.7 km, 64-kt sample HAFS-A; >64.9 km, full sample HAFS-B; >102.3 km, 64-kt sample HAFS-B) for HAFS-A (darker shading, left violin) and HAFS-B (lighter shading, right violin).

the large tilt cases for HAFS-B have negative intensity biases (~−6-kt mean at 96–120 h). The shapes of the violin plots also indicate that there are more samples with intensity biases < −20 kt at 96 and 120 h for the initially large tilt subset.

Although no previous study has documented forecast track biases with respect to shear direction or tilt, NHC forecasters have anecdotally noticed possible directionally oriented biases (P. Papin 2023, personal communication). Figure 8 shows the 36-h track forecast positions for HAFS-A (blue) and HAFS-B (red) relative to the best track and rotated with respect to the midlevel shear vector (850–500 hPa; Figs. 8a–c), deep-layer shear vector (850–200 hPa; Figs. 8d–f), and tilt vector (Figs. 8g–i). For those cases with a small initial tilt magnitude (<10 km), there is very little directional bias with respect to the midlevel shear

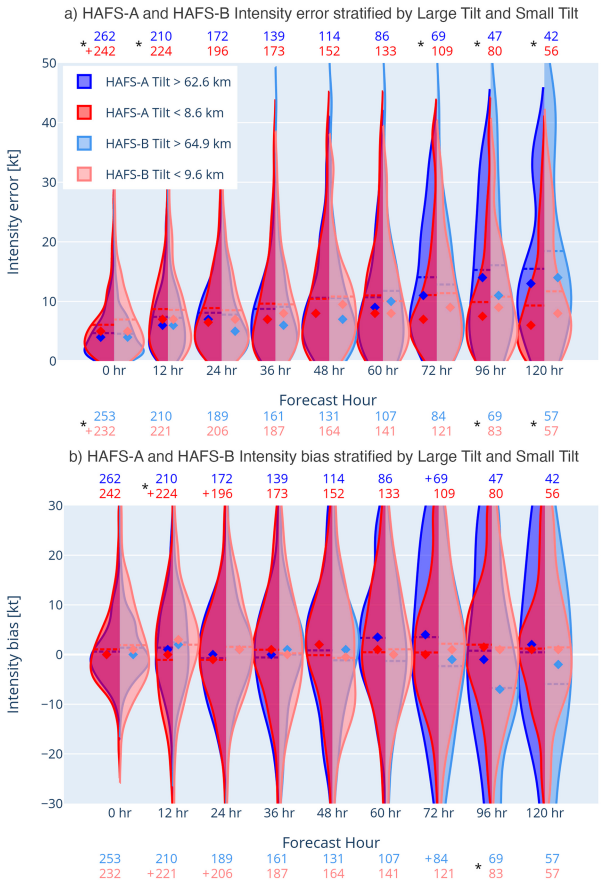

FIG. 7. As in Fig. 4, but for (a) full sample violin plots of intensity error and (b) intensity bias by forecast hour stratified by the lowest 25th percentile of 6-h 2–6.5-km tilt (<8.6 km, HAFS-A; <9.6 km, HAFS-B) and highest 25th percentile (>62.6 km, HAFS-A; >64.9 km, HAFS-B) for HAFS-A (darker shading, left violin) and HAFS-B (lighter shading, right violin).

vector (Fig. 8a; confidence ellipses); however, more model forecast track errors (albeit small) are biased upshear (evidenced by the larger shaded quadrants) than downshear. Although there is slightly more directional bias [oriented downshear left (DSL) to upshear right (USR)] when using the deep-layer shear vector (Fig. 8d), most of the errors are still clustered within 200 km. The tilt-relative perspective (Fig. 8g) for the small tilt cluster has the least amount of directional bias. HAFS-B cases also tend to have a slightly farther uptilt bias when compared to their HAFS-A counterparts.

For the interquartile range of 2–6.5-km tilt magnitudes (10–75 km; Figs. 8b,e,h), both models show an increase in the distribution of track errors oriented directionally compared to the small tilt subsamples. As tilt magnitude increases, both track errors (also shown in Fig. 6a) and the positional correlation of those errors with shear and tilt direction increase. With respect to the midlevel shear vector (Fig. 8b), the major axes of the confidence ellipses are oriented more left of shear to right of shear than downshear (to upshear). The greater left-of-shear HAFS biases are also reflected in the much larger percent occurrence of cases in those quadrants (larger quadrant shadings) than right-of-shear HAFS biases. With respect to the deep-layer shear vector, the track errors for this tilt magnitude subset more closely straddle a DSL to upshear left (USL) orientation (Fig. 8e); however, there is still a large left-of-shear component, which means these model forecasts are more left of shear than reality. Finally, with respect to the tilt vector, the model forecasts still tend to be clustered more left of tilt than reality (Fig. 8h), a result that is slightly more magnified in HAFS-A than in HAFS-B.

The upper quartile of tilt magnitudes (>75 km) exhibits the largest track errors and most noticeable directional correlation with the shear vector (Figs. 8c,f). The largest model forecast errors with respect to the midlevel (Fig. 8c) and deep-layer (Fig. 8f) shear vectors dominate the DSL quadrant. One possible cause of this DSL bias explored later in the section is that the model may produce too much convection downshear in misaligned TCs. This convective bias can lead to new center formation or migration toward that quadrant and a corresponding positive bias in intensity forecasts. The positional relationship of track errors is weaker and less apparent when quantifying with respect to the tilt vector (Fig. 8i), as evidenced by the smaller long-axis (north–south) and short-axis (west–east) confidence ellipse differences.

In addition to quantifying the positional biases with respect to tilt and shear vectors, Fig. 9 adds another dimension by color coding each 36-h forecast with its corresponding intensity bias (kt). Figures 9a and 9b show that negative intensity biases dominate the upshear quadrants (with respect to the deep layer shear vector) with upshear left (USL) median intensity biases of 0 and −1 kt and USR median intensity biases of −4 and −2 kt for track errors < 100 km. DSL and downshear right (DSR) positional biases, on the other hand, tend to skew toward positive intensity biases (+7- and +9-kt median, Fig. 9a). The large initial tilt cases (Fig. 9c) maintain this pattern, and many negative biases also exist in USL, particularly for those with track errors < 100 km (−3-kt median intensity bias); overall, a majority of the larger negative intensity biases (<−10 kt) have track errors less than 100 km. In summary, Fig. 9 reveals the relationship between the larger downshear track errors associated with positive intensity biases and smaller track errors (upshear) associated with negative intensity biases. One hypothesis that will be explored in the next figure is as follows: The TC simulations with DSL track errors and positive intensity biases also have more convection on average (in those quadrants), which results in more center reformation and alignments than observed.

Figure 10 shows shear-relative track errors like Figs. 8 and 9 but with the color-shaded bins indicating the average convective coverage (%) within the inner and outer cores (averages within 0.75 × RMW–2 × RMW for each 3-h output throughout the 0–33 h period preceding the 36-h forecast). The precipitation is partitioned into convection based on a modified algorithm of Steiner et al. (1995) using model reflectivity (Rogers 2010). Figure 10a combines all HAFS-A and HAFS-B cases and shows that those cases with larger downshear track errors tend to have a greater (%) coverage of convection. On the other hand, those with larger upshear (left) track errors

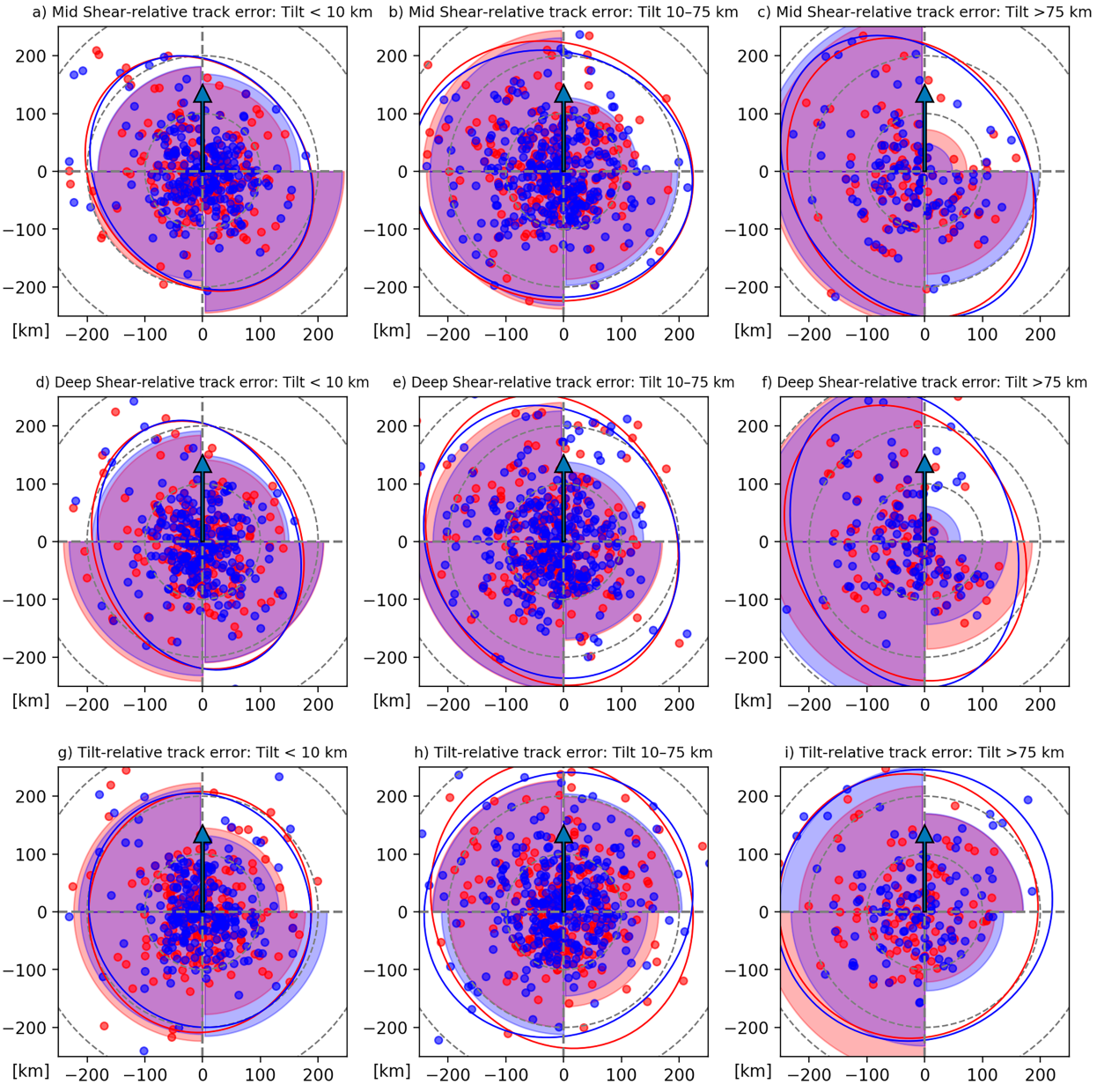

FIG. 8. (a)–(c) The midlevel shear-relative (shear vector pointing up), (d)–(f) deep-layer shear-relative, and (g)–(i) tilt-relative 36-h track errors (km) for all HAFS-A (blue) and HAFS-B (red) cases with (a),(d),(g) tilt < 10 km, (b),(e),(h) 10–75-km tilt, and (e),(f),(i) tilt > 75 km. Confidence ellipses (unshaded, within three standard deviations) are depicted for HAFS-A (blue) and HAFS-B (red). Points are the forecasted position relative to the official track position rotated with respect to the shear or tilt direction (pointing up). Color shadings (blue, HAFS-A; red, HAFS-B) covering each shear or tilt-relative quadrant represent the fraction or percent of all cases that fall within that quadrant with the 200-km dashed circles corresponding to approximately 25%.

tend to have smaller (%) coverages of convection, though there are a few "outlier" pixels (with small sample sizes) that do not follow these trends. The results in Fig. 10a validate the hypothesis that TC simulations with downshear left track biases also have more convection (and speculate that it is overactive) on average than TCs with upshear track biases. Because the model also has upshear (left) track biases (with larger tilt magnitudes, not shown) with less convection (possibly insufficient and less than reality), this signifies a potential deficiency in the model to properly reduce large vortex tilt magnitudes or sustain an alignment of the TC vortex. To more directly link the convective coverages and spatial track errors to intensity biases, Fig. 10b plots only those cases with intensity biases > 10 kt. While the greater (%) convective coverage dominates downshear track errors, upshear track errors less than 50 km also have large convective

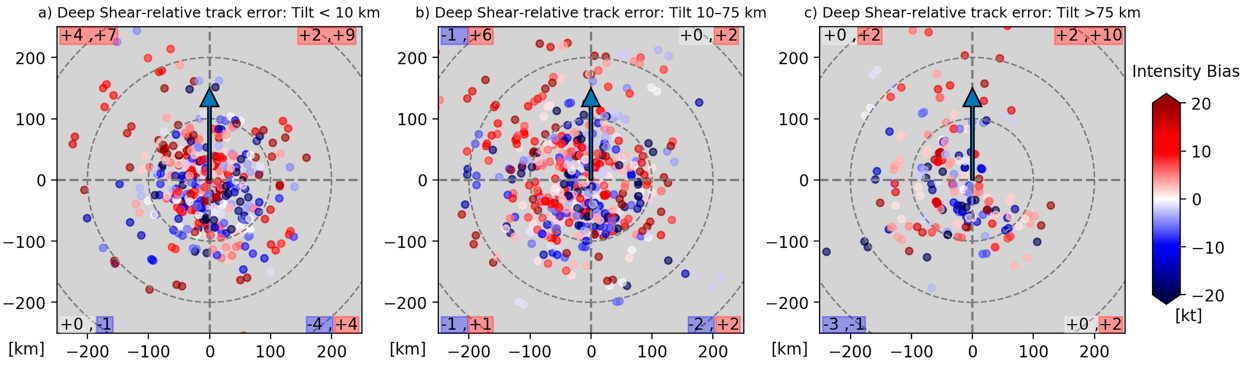

FIG. 9. The deep-layer shear-relative (shear vector pointing up) 36-h track errors (km) for all HAFS-A and HAFS-B cases with (a) vortex tilt < 10 km, (b) 10–75-km tilt, and (c) tilt > 75 km. Points are the forecasted position relative to the official track position rotated with respect to the shear direction (pointing up). Color-shaded circles represent the 36-h intensity errors corresponding to each 36-h spatial track error. The numbers in each shear-relative quadrant indicate that quadrant's median intensity bias (kt) for all track errors < 100 km (first number) and all track errors > 100 km (second number).

coverages. This may potentially be attributed to excessively strong or large eyewalls in TCs with larger positive intensity biases. The higher mid-RH for TCs with positive intensity biases shown previously in Fig. 5 may also be related to the larger convective coverages seen in these cases. Figure 10c stratifies by intensity biases <−10 kt (underforecasted intensification by the model), and most of the larger convective coverages (with larger sample sizes) are cases with smaller track errors (<100 km). Most of the smaller convective coverages occur in cases where the model track error is too far USL. We speculate that HAFS simulations with less convection are more likely to under intensify the TC and have a position error too far USL. Although not demonstrated in this study, the potential effects of tilt changes can also affect spatial biases and intensity errors.

d. *Geographical stratifications of forecast errors and internal storm dynamical parameters*

Different models have specific biases that may result from incorrect initial conditions due to insufficient data assimilated, inadequate representation of atmospheric processes (caused by parameterizations and resolution), or other unknown sources. Interestingly, forecasters have anecdotally noted that different intrabasin geographical locations can have specific model forecast biases that may vary for different models and versions. One example of this is provided by the NHC forecast discussion from Hurricane Delta in 2020 (Blake 2020): "Model guidance has again shifted westward, like the last cycle, and the official forecast is trended in that direction. However, it remains slightly east of the model consensus, due to a notable westward bias (in the Gulf of Mexico) in some of the guidance during this hurricane season." The goal is to better understand model spatial bias so that forecasters are aware of biases prior to the start of each hurricane season and build upon these analyses to develop objective guidance in the future for forecasters throughout the season.

Figure 11 shows the HAFS-A (Fig. 11a) and HAFS-B (Fig. 11b) 72-h track errors (km) binned by geographical position at the initialization (0 h). Both models have large track errors (>225 km) in the far eastern Atlantic that also stretch into parts of the main development region (MDR;

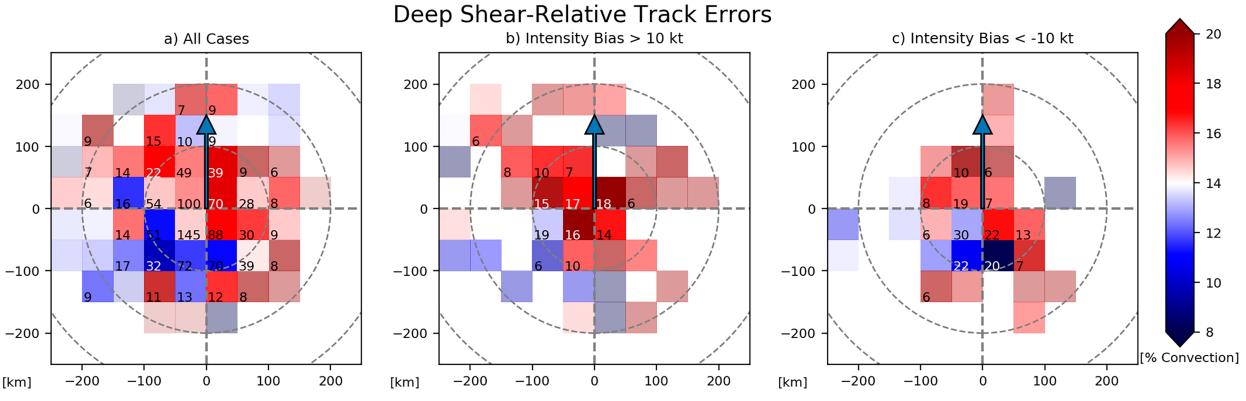

FIG. 10. HAFS-A and HAFS-B deep-layer shear-relative (arrow point up) track errors (km) with color shading indicating the mean convective coverage (% within 0.75 × RMW–2 × RMW for all 3-h outputs throughout the 0–33-h period preceding the 36-h forecast) for each 50-km bin. (a) All cases, (b) stratified by only those cases that have 36-h intensity biases > 10 kt, and (c) those cases with <−10 kt 36-h intensity biases. The numbers denote sample sizes for each bin, and bins without numbers have less than five samples.

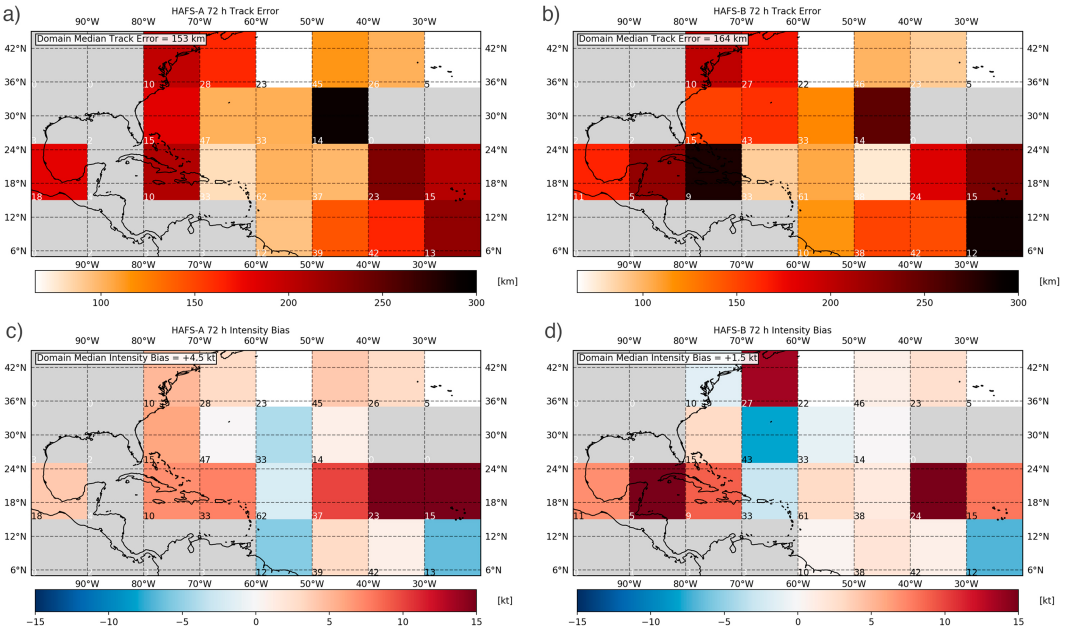

FIG. 11. Atlantic basin retrospective median (a),(b) 72-h forecasted track error (km) and (c),(d) intensity bias (kt) from (a),(c) HAFS-A and (b),(d) HAFS-B averaged across 10° longitude × 10° latitude bins. Each bin represents the 0-h forecast position, and the sample size for each bin is in the lower-left corners. All cases with a 72-h or less forecast over land or verification position over land were removed from the sample, and all bins with less than five samples are shaded in gray. The domain median track errors and intensity biases are displayed for each panel in the upper left.

~10°–20°N, Goldenberg and Shapiro 1996). On the other hand, some of the smallest track forecast errors are found within the central North Atlantic (35°–45°N and 30°–60°W) extending southward into the subtropical Atlantic and northwest Caribbean Sea (15°–35°N and 50°–70°W). Track errors in the southwest Gulf of Mexico, W Caribbean Sea, and W Atlantic are near or above the median (153 km, HAFS-A; 164 km, HAFS-B) in both HAFS-A and HAFS-B.

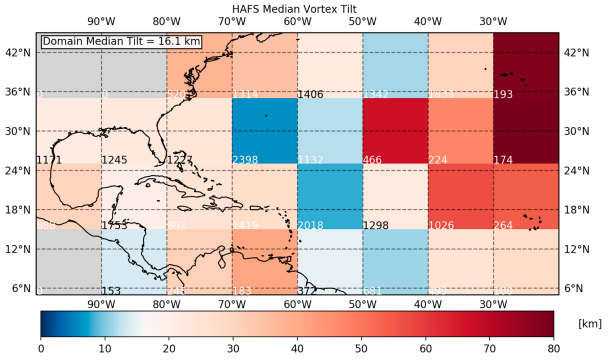

FIG. 12. The median of Atlantic basin retrospective 2–6.5 vortex tilt (km) from HAFS-A and HAFS-B for all forecast hours across 10° longitude × 10° latitude bins. Each bin represents the TC position at the time, and the sample size for each bin is in the lower-left corners. All cases over land were removed from the sample, and all bins with 50 samples or less are shaded in gray. The domain median tilt is displayed in the upper left.

Figures 11c and 11d show similar types of geographical plots but for intensity bias (kt; color shading). Both HAFS-A and HAFS-B again have the largest 72-h positive intensity biases (on average) in the southeastern North Atlantic (and eastern MDR). This can be partly attributed to poor initialization caused by a lack of aircraft reconnaissance and observational data in that part of the Atlantic basin; a similar result has also been observed with HWRF and SHIPS (not shown). In addition, Dunion and Velden (2004) showed an overintensification in SHIPS of systems encountering the Saharan air layer (SAL) in the eastern MDR. Finally, HAFS-A (Fig. 11c) and HAFS-B (Fig. 11d) have positive intensity biases in the northwest–central Caribbean Sea extending northward into the Bahamas, largely the result of overforecasted intensification in moderately sheared TCs like Marco (2020). It is important to note that while some of these spatial biases have also been noted persistently in other operational models like HWRF (not shown), biases can vary significantly among different models due to the dynamical core, physical parameterization (Chen et al. 2023), or initialization differences (Ditchek and Sippel 2023). Furthermore, even the same model over different years can have varying bias distributions, which can also be attributed to storm characteristics that vary from year to year.

In addition to forecast errors, these datasets present a unique opportunity to investigate any potential geographical distributions of vortex tilt in the Atlantic basin for the first time. Figure 12 shows bin-averaged vortex tilt for all HAFS-A and HAFS-B forecast hours. An interesting pattern emerges in which western Atlantic, Gulf of Mexico, and

Caribbean regions all have a median tilt magnitude of 20–30 km (with a large variability; many storms are aligned, and many are also misaligned), which is slightly above the domain median tilt of 16.1 km. The central Atlantic has the lowest median tilt magnitudes (<10 km), which is the same region that features some of the smallest track errors and intensity biases (as shown in Fig. 11). The eastern Atlantic, on the other hand, has some of the largest mean tilt magnitudes (>50 km), which also coincides with weaker TCs on average in this part of the basin and large track and intensity errors seen in Fig. 11. The eastern Atlantic cases also account for many of the large tilt subset TCs with large track errors shown in Fig. 6.

4. Summary

This study uses a 3-yr retrospective sample from the recent operational implementation of HAFS versions A and B to stratify forecast errors by environmental conditions, internal storm dynamic characteristics, and the relationships with intensity (change). Overall, the samples of both HAFS configurations contain a median intensity of tropical storm strength, a reflection of many cases in 2020–22 being characterized by TCs that remained weak until undergoing RI near land (with a relative lack of long track major hurricanes compared to climatology). The sample distributions of HAFS-A and HAFS-B are relatively similar and performed comparably to the average NHC forecast errors for both track and intensity.

Because recent studies have demonstrated the importance of vortex alignment for RI (Frank and Ritchie 2001; Zhang and Tao 2013; Rios-Berrios et al. 2016b, 2018; Chen et al. 2019; Alvey et al. 2020, 2022; Alvey and Hazelton 2022; Stone et al. 2023), these relationships are also explored using a large sample of real case model simulations for the first time to the authors' knowledge. Although very little relationship is observed between vortex tilt and intensity for weak TCs (<64 kt), a more apparent relationship emerges between tilt and future TC intensity change with smaller tilt TCs tending to have greater intensification rates; a large spread in intensity changes still exists for all vortex tilt magnitudes. Nearly all RI cases have initial tilt magnitudes < 40 km, which indicates the importance of a nearly aligned vortex for RI (Alvey et al. 2020, 2022). These results also align with a recent observational analysis of vortex tilt and intensity (change) using a large database of tail Doppler radar data, TC–Radar (Fischer et al. 2023b). Additionally, we find that TCs with larger increases in tilt magnitude tend to coincide with decreases in maximum sustained winds, and TCs with larger decreases in tilt magnitude tend to have larger intensity increases. Finally, the limited potential value of also considering size metrics when evaluating tilt relationships is found, wherein similar relationships to future intensity change are found when investigating tilt/RMW versus the tilt alone. We speculate this could potentially be the result of the RMW definition used (azimuthally averaged 2-km wind with respect to the surface center) or the model's inability to accurately depict the RMW.

TCs within low or moderate VWS (<15–20 kt) and favorable midtropospheric humidity (>50%–60%) regimes tend to have larger intensity errors (negative bias for moderate VWS cases) than systems in high shear. Although Trabing and Bell (2020) were only able to speculate that this was due to RI cases in favorable environments being more directly driven by internal storm dynamics, this study is able to show that HAFS (and likely other high resolution dynamical models) tends to struggle with the critical convective scale and vortex structural processes as evidenced by larger forecast errors associated with larger vortex misalignment. TCs with larger tilt magnitudes tend to have larger forecast track errors and negative intensity biases (at medium range forecasts). Smaller tilt magnitudes, on the other hand, have larger absolute errors for short-range forecasts, which is at least partly attributed to underforecasted intensification in RI cases. The larger tilt cases tend to dominate the far eastern Atlantic, though many still occur within the western Atlantic, wherein the climatological environment is also more conducive for alignment and RI. TCs with a tilted vortex are shown to have both left-of-shear (maximizing DSL) and left-of-tilt-oriented positional track biases. Whereas those cases with larger downshear biases tend to have more convection and larger positive intensity biases, upshear positional biases tend to correspond to less convection and larger negative intensity biases. These relationships highlight the importance of better understanding the interplay of forecast errors and tilt, precipitation, and convection. In addition, because of the larger average errors in tilted TCs and a data assimilation system that primarily leverages a surface position estimate from the NHC to relocate a TC vortex from the previous model 6-h forecast (when inner core data are unavailable), we recommend that future data assimilation model improvements also focus on better capturing the vertical structure of the vortex.

It is important to note that the temporal frequency of variable output (3 h) may not be sufficient to capture all fluctuations in not only traditional metrics like 10-m maximum wind speed (Zhang et al. 2021) but also other internal storm dynamic metrics like vortex tilt and precipitation coverage. While it is certainly plausible that the biases and errors identified in this study may only be attributable to this modeling system, given similar analyses of HWRF (not shown in this study) performed by the authors, it is likely that many of these biases extend across many modeling systems and at least those with similar configurations to HAFS. Future work should aim to incorporate additional modeling systems (including global) and expanded samples across other basins including the east Pacific.

Finally, we summarize a few key recommendations to forecasters, developers, and users of HAFS:

- Internal storm dynamic metrics like vortex tilt may have greater predictive value for future TC intensity and track than traditionally used environmental parameters like vertical wind shear.
- Larger track forecast errors are associated with larger initial vortex misalignment, and larger (short-range) intensity forecast errors are associated with an initially more aligned vortex.

- TCs with a tilted vortex are shown to have directionally oriented positional track biases with respect to both the shear and tilt vectors. On average, the model position forecasts are too far left of shear (and maximize DSL).
- If a HAFS simulation is producing more convection (than the model average), it is more likely to over intensify the TC and has a position error too far downshear.
- Future model improvements should focus on better capturing the (vertical) structure of the vortex through data assimilation improvements and the addition of high-resolution (physics) ensembles.

Acknowledgments. The authors thank Drs. Michael Fischer, John Kaplan (University of Miami and NOAA/OAR/Atlantic Oceanographic and Meteorological Laboratory), and three anonymous reviewers whose comments greatly improved the manuscript. This work was supported by the National Science Foundation under Award AGS-2241605.

Data availability statement. Because of the large file sizes of simulations, these data will be provided only upon request through a remote FTP. Quick view images and plots of the HAFS retrospective data used in this manuscript can be found at https://storm.aoml.noaa.gov/.

REFERENCES

Alland, J. J., B. H. Tang, K. L. Corbosiero, and G. H. Bryan, 2021a: Combined effects of midlevel dry air and vertical wind shear on tropical cyclone development. Part I: Downdraft ventilation. *J. Atmos. Sci.*, **78**, 763–782, https://doi.org/10.1175/JAS-D-20-0054.1.

——, ——, ——, and ——, 2021b: Combined effects of midlevel dry air and vertical wind shear on tropical cyclone development. Part II: Radial ventilation. *J. Atmos. Sci.*, **78**, 783–796, https://doi.org/10.1175/JAS-D-20-0055.1.

Alvey, G. R., III, and Hazelton, A., 2022: How do weak, misaligned tropical cyclones evolve toward alignment? A multicase study using the Hurricane Analysis and Forecast System. *J. Geophys. Res. Atmos.*, **127**, e2022JD037268, https://doi.org/10.1029/2022JD037268.

——, J. Zawislak, and E. Zipser, 2015: Precipitation properties observed during tropical cyclone intensity change. *Mon. Wea. Rev.*, **143**, 4476–4492, https://doi.org/10.1175/MWR-D-15-0065.1.

——, E. Zipser, and J. Zawislak, 2020: How does Hurricane Edouard (2014) evolve toward symmetry before rapid intensification? A high-resolution ensemble study. *J. Atmos. Sci.*, **77**, 1329–1351, https://doi.org/10.1175/JAS-D-18-0355.1.

——, M. Fischer, P. Reasor, J. Zawislak, and R. Rogers, 2022: Observed processes underlying the favorable vortex repositioning early in the development of Hurricane Dorian (2019). *Mon. Wea. Rev.*, **150**, 193–213, https://doi.org/10.1175/MWR-D-21-0069.1.

Bhatia, K. T., and D. S. Nolan, 2013: Relating the skill of tropical cyclone intensity forecasts to the synoptic environment. *Wea. Forecasting*, **28**, 961–980, https://doi.org/10.1175/WAF-D-12-00110.1.

Blake, E., 2020: Hurricane Delta discussion number 7, NWS NHC Forecast Discussion AL262020. NWS NHC, https://www.nhc.noaa.gov/archive/2020/al26/al262020.discus.007.shtml?text.

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.

Chen, H., and S. G. Gopalakrishnan, 2015: A study on the asymmetric rapid intensification of Hurricane Earl (2010) using the HWRF system. *J. Atmos. Sci.*, **72**, 531–550, https://doi.org/10.1175/JAS-D-14-0097.1.

Chen, J.-H., and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

Chen, X., J. A. Zhang, and F. D. Marks, 2019: A thermodynamic pathway leading to rapid intensification of tropical cyclones in shear. *Geophys. Res. Lett.*, **46**, 9241–9251, https://doi.org/10.1029/2019GL083667.

——, G. H. Bryan, A. Hazelton, F. D. Marks, and P. Fitzpatrick, 2022: Evaluation and improvement of a TKE-based eddy-diffusivity mass-flux (EDMF) planetary boundary layer scheme in hurricane conditions. *Wea. Forecasting*, **37**, 935–951, https://doi.org/10.1175/WAF-D-21-0168.1.

——, A. Hazelton, F. D. Marks, G. J. Alaka Jr., and C. Zhang, 2023: Performance of an Improved TKE-based eddy-diffusivity mass-flux (EDMF) PBL scheme in 2021 hurricane forecasts from the Hurricane Analysis and Forecast System. *Wea. Forecasting*, **38**, 321–336, https://doi.org/10.1175/WAF-D-22-0140.1.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, https://doi.org/10.1175/WAF862.1.

DesRosiers, A. J., M. M. Bell, P. J. Klotzbach, M. S. Fischer, and P. D. Reasor, 2023: Observed relationships between tropical cyclone vortex height, intensity, and intensification rate. *Geophys. Res. Lett.*, **50**, e2022GL101877, https://doi.org/10.1029/2022GL101877.

Ditchek, S. D., and J. A. Sippel, 2023: A comparison of the impacts of inner-core, in-vortex, and environmental dropsondes on tropical cyclone forecasts during the 2017–20 hurricane seasons. *Wea. Forecasting*, **38**, 2169–2187, https://doi.org/10.1175/WAF-D-23-0055.1.

Dunion, J. P., and C. S. Velden, 2004: The impact of the Saharan air layer on Atlantic tropical cyclone activity. *Bull. Amer. Meteor. Soc.*, **85**, 353–366, https://doi.org/10.1175/BAMS-85-3-353.

Finocchio, P. M., and S. J. Majumdar, 2017: A statistical perspective on wind profiles and vertical wind shear in tropical cyclone environments of the Northern Hemisphere. *Mon. Wea. Rev.*, **145**, 361–378, https://doi.org/10.1175/MWR-D-16-0221.1.

Fischer, M. S., P. D. Reasor, B. H. Tang, K. L. Corbosiero, R. D. Torn, and X. Chen, 2023a: A tale of two vortex evolutions: Using a high-resolution ensemble to assess the impacts of ventilation on a tropical cyclone rapid intensification event. *Mon. Wea. Rev.*, **151**, 297–320, https://doi.org/10.1175/MWR-D-22-0037.1.

——, R. F. Rogers, P. D. Reasor, and J. P. Dunion, 2023b: An observational analysis of the relationship between tropical cyclone vortex tilt, precipitation structure, and intensity change. *Mon. Wea. Rev.*, **152**, 203–225, https://doi.org/10.1175/MWR-D-23-0089.1.

Frank, W. M., and E. A. Ritchie, 2001: Effects of vertical wind shear on the intensity and structure of numerically simulated

hurricanes. *Mon. Wea. Rev.*, **129**, 2249–2269, https://doi.org/10.1175/1520-0493(2001)129<2249:EOVWSO>2.0.CO;2.

Goldenberg, S. B., and L. J. Shapiro, 1996: Physical mechanisms for the association of El Niño and West African rainfall with Atlantic major hurricane activity. *J. Climate*, **9**, 1169–1187, https://doi.org/10.1175/1520-0442(1996)009<1169:PMFTAO>2.0.CO;2.

Guimond, S. R., G. M. Heymsfield, P. D. Reasor, and A. C. Didlake Jr., 2016: The rapid intensification of Hurricane Karl (2010): New remote sensing observations of convective bursts from the Global Hawk platform. *J. Atmos. Sci.*, **73**, 3617–3639, https://doi.org/10.1175/JAS-D-16-0026.1.

Hazelton, A., G. J. Alaka Jr., L. Cowan, M. Fischer, and S. Gopalakrishnan, 2021: Understanding the processes causing the early intensification of Hurricane Dorian through an ensemble of the Hurricane Analysis and Forecast System (HAFS). *Atmosphere*, **12**, 93, https://doi.org/10.3390/atmos12010093.

——, and Coauthors, 2023: 2022 real-time hurricane forecasts from an experimental version of the Hurricane Analysis and Forecast System (HAFSV0.3S). *Front. Earth Sci.*, **11**, 1264969, https://doi.org/10.3389/feart.2023.1264969.

Jiang, H., 2012: The relationship between tropical cyclone intensity change and the strength of inner-core convection. *Mon. Wea. Rev.*, **140**, 1164–1176, https://doi.org/10.1175/MWR-D-11-00134.1.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

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.

Leighton, H., S. Gopalakrishnan, J. A. Zhang, R. F. Rogers, Z. Zhang, and V. Tallapragada, 2018: Azimuthal distribution of deep convection, environmental factors, and tropical cyclone rapid intensification: A perspective from HWRF ensemble forecasts of Hurricane Edouard (2014). *J. Atmos. Sci.*, **75**, 275–295, https://doi.org/10.1175/JAS-D-17-0171.1.

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

Munsell, E. B., F. Zhang, and D. P. Stern, 2013: Predictability and dynamics of a nonintensifying tropical storm: Erika (2009). *J. Atmos. Sci.*, **70**, 2505–2524, https://doi.org/10.1175/JAS-D-12-0243.1.

——, ——, J. A. Sippel, S. A. Braun, and Y. Weng, 2017: Dynamics and predictability of the intensification of Hurricane Edouard (2014). *J. Atmos. Sci.*, **74**, 573–595, https://doi.org/10.1175/JAS-D-16-0018.1.

Nam, C. C., M. M. Bell, and D. Tao, 2023: Bifurcation points for tropical cyclone genesis and intensification in sheared and dry environments. *J. Atmos. Sci.*, **80**, 2239–2259, https://doi.org/10.1175/JAS-D-22-0100.1.

Nguyen, L. T., and J. Molinari, 2015: Simulation of the downshear reformation of a tropical cyclone. *J. Atmos. Sci.*, **72**, 4529–4551, https://doi.org/10.1175/JAS-D-15-0036.1.

——, ——, and D. Thomas, 2014: Evaluation of tropical cyclone center identification methods in numerical models. *Mon. Wea. Rev.*, **142**, 4326–4339, https://doi.org/10.1175/MWR-D-14-00044.1.

Rappin, E. D., and D. S. Nolan, 2012: The effect of vertical shear orientation on tropical cyclogenesis. *Quart. J. Roy. Meteor. Soc.*, **138**, 1035–1054, https://doi.org/10.1002/qj.977.

Reasor, P. D., and M. D. Eastin, 2012: Rapidly intensifying Hurricane Guillermo (1997). Part II: Resilience in shear. *Mon. Wea. Rev.*, **140**, 425–444, https://doi.org/10.1175/MWR-D-11-00080.1.

Rios-Berrios, R., 2020: Impacts of radiation and cold pools on the intensity and vortex tilt of weak tropical cyclones interacting with vertical wind shear. *J. Atmos. Sci.*, **77**, 669–689, https://doi.org/10.1175/JAS-D-19-0159.1.

——, and R. D. Torn, 2017: Climatological analysis of tropical cyclone intensity changes under moderate vertical wind shear. *Mon. Wea. Rev.*, **145**, 1717–1738, https://doi.org/10.1175/MWR-D-16-0350.1.

——, ——, and C. A. Davis, 2016a: An ensemble approach to investigate tropical cyclone intensification in sheared environments. Part I: Katia (2011). *J. Atmos. Sci.*, **73**, 71–93, https://doi.org/10.1175/JAS-D-15-0052.1.

——, ——, and ——, 2016b: An ensemble approach to investigate tropical cyclone intensification in sheared environments. Part II: Ophelia (2011). *J. Atmos. Sci.*, **73**, 1555–1575, https://doi.org/10.1175/JAS-D-15-0245.1.

——, C. A. Davis, and R. D. Torn, 2018: A hypothesis for the intensification of tropical cyclones under moderate vertical wind shear. *J. Atmos. Sci.*, **75**, 4149–4173, https://doi.org/10.1175/JAS-D-18-0070.1.

Rogers, R., 2010: Convective-scale structure and evolution during a high-resolution simulation of tropical cyclone rapid intensification. *J. Atmos. Sci.*, **67**, 44–70, https://doi.org/10.1175/2009JAS3122.1.

——, P. D. Reasor, J. A. Zawislak, and L. T. Nguyen, 2020: Precipitation processes and vortex alignment during the intensification of a weak tropical cyclone in moderate vertical shear. *Mon. Wea. Rev.*, **148**, 1899–1929, https://doi.org/10.1175/MWR-D-19-0315.1.

Schecter, D. A., 2022: Intensification of tilted tropical cyclones over relatively cool and warm oceans in idealized numerical simulations. *J. Atmos. Sci.*, **79**, 485–512, https://doi.org/10.1175/JAS-D-21-0051.1.

——, and K. Menelaou, 2020: Development of a misaligned tropical cyclone. *J. Atmos. Sci.*, **77**, 79–111, https://doi.org/10.1175/JAS-D-19-0074.1.

Steiner, M., R. A. Houze Jr., and S. E. Yuter, 1995: Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. *J. Appl. Meteor.*, **34**, 1978–2007, https://doi.org/10.1175/1520-0450(1995)034<1978:CCOTDS>2.0.CO;2.

Stone, Ž., G. R. Alvey III, J. P. Dunion, M. S. Fischer, D. J. Raymond, R. F. Rogers, S. Sentić, and J. Zawislak, 2023: Thermodynamic contribution to vortex alignment and rapid intensification of Hurricane Sally (2020). *Mon. Wea. Rev.*, **151**, 931–951, https://doi.org/10.1175/MWR-D-22-0201.1.

Sumwalt, R. L., III, C. A. Hart, E. F. Weener, and T. B. Dinh-Zahr, 2017: Tropical cyclone information for mariners. National Transportation Safety Board Safety Recommendation Rep. DCA16MM001, 21 pp.

Tang, B., and K. Emanuel, 2010: Midlevel ventilation's constraint on tropical cyclone intensity. *J. Atmos. Sci.*, **67**, 1817–1830, https://doi.org/10.1175/2010JAS3318.1.

——, and ——, 2012: A ventilation index for tropical cyclones. *Bull. Amer. Meteor. Soc.*, **93**, 1901–1912, https://doi.org/10.1175/BAMS-D-11-00165.1.

Tao, C., and H. Jiang, 2015: Distributions of shallow to very deep precipitation–convection in rapidly intensifying tropical cyclones. *J. Climate*, **28**, 8791–8824, https://doi.org/10.1175/JCLI-D-14-00448.1.

Tao, D., and F. Zhang, 2014: Effect of environmental shear, sea-surface temperature, and ambient moisture on the formation and predictability of tropical cyclones: An ensemble-mean perspective. *J. Adv. Model. Earth Syst.*, **6**, 384–404, https://doi.org/10.1002/2014MS000314.

——, and ——, 2015: Effects of vertical wind shear on the predictability of tropical cyclones: Practical versus intrinsic limit. *J. Adv. Model. Earth Syst.*, **7**, 1534–1553, https://doi.org/10.1002/2015MS000474.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Trabing, B. C., and M. M. Bell, 2020: Understanding error distributions of hurricane intensity forecasts during rapid intensity changes. *Wea. Forecasting*, **35**, 2219–2234, https://doi.org/10.1175/WAF-D-19-0253.1.

Yu, C.-L., B. Tang, and R. G. Fovell, 2023: Tropical cyclone tilt and precession in moderate shear: Precession hiatus in a critical shear regime. *J. Atmos. Sci.*, **80**, 909–932, https://doi.org/10.1175/JAS-D-22-0200.1.

Zawislak, J., H. Jiang, G. R. Alvey III, E. J. Zipser, R. F. Rogers, J. A. Zhang, and S. N. Stevenson, 2016: Observations of the structure and evolution of Hurricane Edouard (2014) during intensity change. Part I: Relationship between the thermodynamic structure and precipitation. *Mon. Wea. Rev.*, **144**, 3333–3354, https://doi.org/10.1175/MWR-D-16-0018.1.

Zhang, F., and D. Tao, 2013: Effects of vertical wind shear on the predictability of tropical cyclones. *J. Atmos. Sci.*, **70**, 975–983, https://doi.org/10.1175/JAS-D-12-0133.1.

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.

Zhou, L., and Coauthors, 2022: Improving global weather prediction in GFDL SHiELD through an upgraded GFDL cloud microphysics scheme. *J. Adv. Model. Earth Syst.*, **14**, e2021MS002971, https://doi.org/10.1029/2021MS002971.