Evaluation of Experimental High-Resolution Model Forecasts of Tropical Cyclone Precipitation using Object-Based Metrics

Shakira D. Stackhouse,[a] Stephanie E. Zick,[a] Corene J. Matyas,[b] Kimberly M. Wood,[c] Andrew T. Hazelton,[d,e] and Ghassan J. Alaka Jr.[d]

[a] *Department of Geography, Virginia Tech, Blacksburg, VA*

[b] *Department of Geography, University of Florida, Gainesville, FL*

[c] *Department of Geosciences, Mississippi State University, Mississippi State University, MS*

[d] *NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, FL, USA*

[e] *University of Miami/Cooperative Institute for Marine and Atmospheric Studies, Miami, FL, USA*

Corresponding author: Stephanie Zick, sezick@vt.edu

ABSTRACT

Tropical cyclone (TC) precipitation poses serious hazards including freshwater flooding. High-resolution hurricane models predict the location and intensity of TC rainfall, which can influence local evacuation and preparedness policies. This study evaluates 0–72-hour precipitation forecasts from two experimental models, the Hurricane Analysis and Forecast System (HAFS) model and the Basin-scale Hurricane Weather Research and Forecasting (HWRF-B) model, for 2020 North Atlantic landfalling TCs. We use an object-based method that quantifies the shape and size of the forecast and observed precipitation. Precipitation objects are then compared for light, moderate, and heavy precipitation using spatial metrics (e.g., area, perimeter, elongation). Results show that both models forecast precipitation that is too connected, too close to the TC center, too enclosed around the TC center. Collectively, these spatial biases suggest that the model forecasts are too intense even though there is a negative intensity bias for both models, indicating there may be an inconsistency between the precipitation configuration and the maximum sustained winds in the model forecasts. The HAFS model struggles with forecasting stratiform versus convective precipitation and with the representation of lighter (stratiform) precipitation during the first six hours after initialization. No such spin-up issues are seen in the HWRF-B forecasts, which instead exhibit systematic biases at all lead times and systematic issues across all rain rate thresholds. Future work will investigate spin-up issues in the HAFS model forecast and how the microphysics parameterization affects the representation of precipitation in both models.

1. Introduction

Forecast verification is used to determine the skillfulness of numerical weather prediction (NWP) models at forecasting weather-related variables. For example, precipitation verification is performed due to the extreme impacts it can have on humans and the environment. The most significant hazard associated with precipitation is flooding, which can result in many fatalities, extensive structural damage, and negative economic repercussions. According to Rappaport (2014), rainfall-induced flooding was the second leading cause of death behind storm surge for tropical cyclones (TCs) that occurred from 1963 to 2012. Due to the dangers of rainfall, affected areas must be able to adequately prepare for landfalling TCs. Accurate forecasts from hurricane models greatly assist in this preparation.

This research evaluates precipitation forecast skill for two experimental National Oceanic and Atmospheric Administration (NOAA) hurricane models: the Hurricane Analysis and Forecast System (HAFS) and the Basin-Scale Hurricane Weather Research and Forecasting model (HWRF-B hereafter). The HAFS model is an application of the Unified Forecast System (UFS), a community-based modeling system at NOAA that provides enhancements to operational forecasting (Unified Forecast System, 2022). There are two versions of the HAFS model: the stand-alone regional domain configuration (HAFS-SAR) and the global-nesting configuration (HAFS-globalnest). The HAFS-SAR (Dong et al. 2020) is a limited area model (LAM) that contains a high-resolution static nest centered about the North Atlantic region, while the HAFS-globalnest features a global domain in addition to a regional nest over the North Atlantic region in which there is two-way feedback (Hazelton et al. 2021a, 2022a). This study evaluates the HAFS-globalnest; hereafter, the model will be referred to as HAFS. The HWRF-B model was developed at the NOAA/Atlantic Oceanographic and Meteorological Laboratory (AOML) Hurricane Research Division (HRD) to address shortcomings in operational HWRF related to multi-scale and multi-storm interactions (Zhang et al. 2016; Alaka et al. 2017, 2020, 2022). The HWRF-B has a large static outermost domain and contains high-resolution movable multi-level nests that follow multiple tropical systems of interest. Conversely, the operational HWRF only accommodates nests for a single TC per forecast integration. The outer domain of the HWRF-B produces forecasts of the large-scale environment while the two telescopic nests produce high-resolution forecasts for each tropical system of interest, including sharp gradients in the inner core region (Alaka et al. 2022).

Numerical approaches implemented to verify precipitation forecasts fall into two general categories: point-based and object-based methods. Point-based methods are more traditional and compare the pixel values from forecast and observed grids using a variety of statistics (e.g., Pearson's correlation) (Wilks 2011). Object-based methods are executed by first converting the forecast and observed precipitation fields into objects through a rain rate thresholding process. Spatial metrics are then calculated and compared for the forecast and observed objects (see section 2). Although point-based methods provide a general assessment of the forecast skill, a major drawback is the "double penalty" that occurs when precipitation is forecast correctly but in the wrong location (Davis et al. 2006; Rossa et al. 2008; Gilleland et al. 2010). This penalizes the forecast twice: first for failing to forecast the rainfall in the correct location and second for generating a false alarm in an incorrect location. This double penalty issue has become more pronounced with high-resolution forecasts that can produce smaller-scale features that are slightly displaced in space and/or time. In this study, we first present point-based methods and object-based methods in two case studies to illustrate the two approaches and the advantages of an object-based approach.

One of the earliest studies to implement an object-based method to evaluate precipitation forecasts was Ebert and McBride (2000) in which they used contiguous rain areas (CRAs; unified, rain rate threshold-defined forecast and observed entities that overlap or are near each other) to determine displacement and intensity errors. This technique involved pattern matching where the forecast field was translated over the observed field until a best fit condition was satisfied. Davis et al. (2006) developed an object-based approach similar to the CRA technique where the rain area identification process included thresholding and filtering procedures to result in whole objects with smoothed boundaries. The forecast and observed objects were then matched depending on the separation distance between their centroids. Errors between matched objects were computed and compared for the centroid position, rainfall intensity, area, and other spatial attributes (Davis et al. 2006). Li et al. (2016) used a unique approach to define precipitation objects through a watershed transformation. Precipitation intensities were depicted as altitudes on a topographic surface, which were represented by pixels. The maximum pixel values were treated as storm centers and other precipitation pixels were assigned to the storm center closest to them in terms of topographic distance, which created the objects. The forecast and observed precipitation objects were matched based on whether they overlapped, and an overall score was determined depending on the values of various geometric measures.

Zick and Matyas (2016) demonstrated an object-based method that quantified TC precipitation fields using shape metrics. The objects were created based on the satisfaction of search radius and rain rate threshold conditions. Shape metrics, which included asymmetry, fragmentation, and dispersiveness, were calculated across all objects at different timesteps for Hurricane Katrina. A statistical test was used to identify significant differences between the shape metric values throughout the TC lifecycle, which helped in identifying timesteps when the precipitation was rapidly evolving. A similar approach can be used to identify significant differences between forecast and observed shape metric values (Matyas et al. 2018; Zick 2020; Zick et al. 2022). This study applies a similar object-based approach to evaluate the HAFS and HWRF-B models versus observations.

Model verification is relevant to both model developers and forecasters. Developers often seek opportunities to enhance models. Thus, identification of the strengths and weaknesses of hurricane models can show where improvements can be made. The skillfulness of TC precipitation forecasts is also useful information for forecasters because it allows them to better understand the aspects for which the model succeeds and fails. Forecasts can then be adjusted depending on the findings from the model evaluation (i.e., if the HAFS tends to "under-forecast" the extent of TC rainfall, a forecaster might forecast precipitation for a wider area than what is indicated by the model).

Since the HAFS and HWRF-B models were both experimental (i.e., non-operational) during the 2020 hurricane season, only a few studies have examined their performance. TC track and intensity are usually prioritized for model verification, so there is a general lack of TC rainfall evaluation studies. Improving the skillfulness of TC precipitation forecasts in numerical models should be a focal point because of the hazards associated with torrential rainfall. Also, an increase in the accuracy of TC precipitation forecasts may support improvements in track and intensity forecasts due to the relationship between TC cloud features and TC intensity indicated by the Dvorak Technique (Dvorak 1990; Velden et al. 2006). This study will provide an in-depth analysis of the HAFS and HWRF-B rainfall forecasts of landfalling TCs.

The following research questions are investigated in this study: 1) How do point- and object-based precipitation verification methods compare with each other for evaluating high-resolution TC forecasts? 2) How accurate are the HAFS and HWRF-B models at predicting the short-range rainfall distribution in tropical cyclones? and 3) How do the verification metrics of the HAFS and HWRF-B precipitation forecasts vary by lead time? Each model is

evaluated individually versus the observations due to slightly different samples for each model. Due to point-based methods tending to over-penalize forecasts when there are location errors (Davis et al. 2006; Rossa et al. 2008; Gilleland et al. 2010), the object-based methods are expected to more closely represent the accuracy of the hurricane models. In an evaluation of the HAFS skill in forecasting TCs in (Hazelton et al. 2021a), the model tended to have a negative bias in forecasting the intensity of the storms. However, for the 2020 season, the HAFS model overpredicted intensity due to a lack of ocean coupling (Hazelton et al. 2022a). In this study, we calculate the track and intensity biases associated with the landfalling storms. We expect the intensity bias to be reflected in the spatial characteristics of the TC precipitation (e.g., that a positive intensity bias would be associated with a more enclosed eyewall). Both model forecasts may be negatively impacted by model spin-up during earlier lead times when dynamical and physical adjustments in the models are required to reach a balanced state (Sun et al. 2014; Wong and Skamarock 2016; Chu et al. 2018), although spin-up issues should be less prevalent in HWRF-B due to its vortex initialization scheme (Biswas et al. 2018; Liu et al. 2020). In addition, skill performance may degrade with time as commonly observed in model forecasts (Lorenz 1963; Hohenegger and Schar 2007).

The paper is organized as follows: information about data is provided in section 2, methods are provided in section 3, point-based metrics are presented in section 4, case studies of Hurricanes Isaias and Laura are investigated in section 5, all landfalling storms from the 2020 North Atlantic hurricane season are evaluated in section 6, and conclusions are presented in section 7.

2. Data

a. Stage IV

The National Centers for Environmental Prediction (NCEP) Stage IV analysis data, produced by the National Weather Service (NWS) River Forecast Centers (RFC), serves as the observational dataset used to verify the TC precipitation forecasts from the HAFS model. Stage IV consists of precipitation data retrieved from NOAA's Next Generation Weather Radars (NEXRAD) system with a rain gauge correction and is available as hourly, 6-hourly, and 24-hourly accumulations (Lin and Mitchell 2005; NOAA 2020). The grid spacing of Stage IV is approximately 4 km and its domain covers the contiguous United States (Lin and Mitchell 2005).

Stage IV is commonly used for model verification studies centered on precipitation forecasts (Davis et al. 2006; Luitel et al. 2016; Li et al. 2016; Zick 2020). Additionally, this dataset has been shown to be useful in estimating convective precipitation (Nelson et al. 2015). One disadvantage of using Stage IV is the inaccuracies that may occur due to the automated process implemented for the quality control of hourly data (Nelson et al. 2015). Individual radars can also go down in extreme weather events, as seen in Hurricane Laura (see section 4), which can lead to Stage IV precipitation underestimates in the vicinity of that radar.

b. Models

This study investigates precipitation forecasts in two models: the global-nested version of HAFS and the HWRF-B. The model specifications are summarized here; further details on the HAFS and HWRF-B models are available in Hazelton et al. (2022a) and Alaka et al. (2020, 2022), respectively. The HAFS model has a 13-km global domain along with a static nest over the North Atlantic region. The static nest has a horizontal resolution of 3 km, and 75 vertical levels are utilized for the global and regional domains. For the dynamical core, the HAFS implements the nested version of the finite-volume cubed-sphere (FV3) (Harris and Lin 2013). In a study by Hazelton et al. (2018), the nested FV3 provided promising results in forecasting TC structure, shown by its good performance in forecasting the maximum wind radii. Vortex initialization and ocean coupling were not yet implemented in the experimental version of HAFS used in this study, although an updated version includes those capabilities. For the physics parameterizations, many of the schemes are shared between the HAFS and the operational GFS, but some, such as the planetary boundary layer (PBL) scheme, have been modified to better suit the environment around TCs (Hazelton et al. 2021b, 2022b).

The basin-scale (i.e., multi-storm) version of the HWRF model (i.e., HWRF-B) is also assessed. HWRF is a state-of-the-art numerical hurricane prediction system that has been operational at NCEP since 2007; it has improved TC intensity forecasts through advanced data assimilation procedures, more appropriate physics parameterizations, and other upgrades (Gopalakrishnan et.al., 2021). HWRF is a triply-nested modeling system, with an outermost domain that has a horizontal resolution of 13.5 km and two telescopic moving nest domains with horizontal resolutions of 4.5 km and 1.5 km, respectively, that follow a tropical system of interest. There are 75 vertical pressure-sigma levels and a model top of 10 hPa. HWRF utilizes the Nonhydrostatic Mesoscale Model (NMM) for its atmospheric dynamic core, and

198 the atmosphere is coupled to the Princeton Ocean Model (POM). HWRF uses a vortex
199 initialization scheme developed by Kurihara et al. (1995) that has been shown to reduce
200 model spin-down issues (Liu et al. 2020). For more details about the operational HWRF
201 modeling system, including details about physics parameterizations, please refer to the
202 HWRF documentation (Biswas et al. 2018).

203 The HWRF-B model is identical to the operational HWRF except for two advanced
204 configuration options (Zhang et al. 2016; Alaka et al. 2017, 2020, 2022). For one, the
205 outermost domain in HWRF-B is much larger than in the operational HWRF, spanning the
206 entire National Hurricane Center (NHC) area of responsibility (i.e., North Atlantic and
207 eastern North Pacific basins), and, unlike the operational HWRF, the HWRF-B outermost
208 domain is not recentered on the TC at the forecast initialization time. The second difference is
209 that HWRF-B can be configured with moving nests for multiple tropical systems (up to five
210 in the 2020 version). Alaka et al. (2022) showed that HWRF-B improved track and intensity
211 forecasts compared to the operational HWRF, which motivated its use in this study. Like the
212 operational HWRF, the innermost nest in HWRF-B has a grid spacing of approximately 1.5-
213 km.

214 This research evaluates the forecast precipitation for nine North Atlantic TCs of at least
215 tropical storm intensity that made landfall in the U.S. in 2020 (Table 1). This study focuses
216 on landfalling TCs because the Stage IV domain is limited to the contiguous United States.
217 Three landfalling storms are excluded (Tropical Storm Bertha, Tropical Storm Cristobal, and
218 Hurricane Marco) due to either the lack of HAFS model data or Stage IV coverage.
219 Additionally, Tropical Storm Fay and Hurricane Eta are only included in the lead time
220 analysis (section 6b) because of their short study periods that do not allow for sufficient time
221 steps for evaluating the storms individually (section 6a). Out of the nine storms included in
222 the analysis, two were tropical storms, four were hurricanes, and three were major hurricanes
223 at peak intensity (Table 1).

224 For the individual TCs, increments of 3-hourly accumulated precipitation within the first
225 72 hours of each model simulation for the storm are included in the analysis, after which the
226 forecasts are expected to differ substantially from the observations. Therefore, up to 24 (3-
227 hour) forecasts of each model simulation are evaluated. Four model simulations are run each
228 day as the models are initialized every six hours at 00, 06, 12 and 18 UTC. Table 1 lists the
229 number of model timesteps used in the analysis for each storm. These numbers are dependent
230 on the period of study and the availability of the data. Due to the desire to retain as much data

as possible for analysis in sections 4-6, the two models include slightly different samples (Table 1). This study does not aim to compare the model performance of the two models; instead, we aim to compare each model individually with the observations, which justifies the heterogenous sample. The study period for each TC begins approximately 3-6 hours before landfall and ends around the time when the storm no longer has TC characteristics as designated by the NHC hurricane database (HURDAT2) (Landsea and Franklin 2013). The first model run included in the assessment for each storm is about 48 hours prior to the beginning of the study period, which allows for sufficient timesteps to be evaluated for each storm to ensure an extensive analysis of the HAFS and HWRF-B model performance. Corresponding track and intensity errors for this landfalling TC sample are shown in Figure 1 (for all storms combined) and Table 2 (for individual storms). Overall, HAFS and HWRF-B have similar intensity error statistics for this sample, and HAFS has slightly lower track errors at longer lead times.

Storm Name	Period of Study	Earliest Model Initialization	Total Observed Analyses	Total HAFS Forecast Times	Total HWRF-B Forecast Times
Fay* (TS)	18 UTC 10 Jul - 06 UTC 11 Jul	18 UTC 08 Jul	4	10	34
Hanna (H)	21 UTC 25 Jul - 00 UTC 27 Jul	00 UTC 24 Jul	9	77	66
Isaias (H)	21 UTC 03 Aug - 00 UTC 05 Aug	18 UTC 01 Aug	9	88	79
Laura (MH)	06 UTC 27 Aug - 06 UTC 29 Aug	00 UTC 25 Aug	16	158	162
Sally (H)	09 UTC 16 Sep - 12 UTC 17 Sep	06 UTC 14 Sep	9	86	90
Beta	00 UTC 22 Sep -	06 UTC 20 Sep	8	66	74

(TS)	00 UTC 23 Sep				
Delta (MH)	21 UTC 09 Oct - 18 UTC 10 Oct	18 UTC 07 Oct	7	68	66
Zeta (H)	21 UTC 28 Oct - 18 UTC 29 Oct	18 UTC 26 Oct	7	75	75
Eta* (MH)	21 UTC 8 Nov - 06 UTC 9 Nov	12 UTC 7 Nov	3	18	26

Table 1. Details of U.S. landfalling Atlantic TCs in 2020 that are included in the study. Storms with an asterisk (*) are only included in the lead time analysis. Parenthetical letters indicate peak intensity over the storm lifecycle (TS = Tropical Storm, H = Hurricane, MH = Major Hurricane).

	Sample size		Mean absolute track error (km)		Mean absolute intensity error (kt)		Mean intensity error/bias (kt)	
	HAFS	HWRF-B	HAFS	HWRF-B	HAFS	HWRF-B	HAFS	HWRF-B
Hanna (AL082020)	24	21	66.3	80.1	8.4	8.3	-3.2	-1.1
Isaias (AL092020)	15	39	89.3	102.9	6.7	9.7	-5.4	-8.4
Laura (AL132020)	86	81	84.8	81.2	7.1	8.0	-3.3	-3.4
Sally (AL192020)	48	43	59.3	79.6	11.8	11.5	-10.8	-11.1
Beta (AL222020)	39	39	51.4	71.6	5.6	6.9	-4.8	-6.7
Delta (AL262020)	34	33	61.6	61.9	4.2	4.3	-1.0	-1.1
Zeta (AL282020)	40	42	51.4	70.1	10.4	8.3	-8.9	-5.2

Table 2. Time-averaged verification of 6-72-h HAFS and HWRF-B forecast errors for landfalling TCs included in this study. Sample sizes for individual storms are also provided for each model. Errors are computed relative to the HURDAT2.

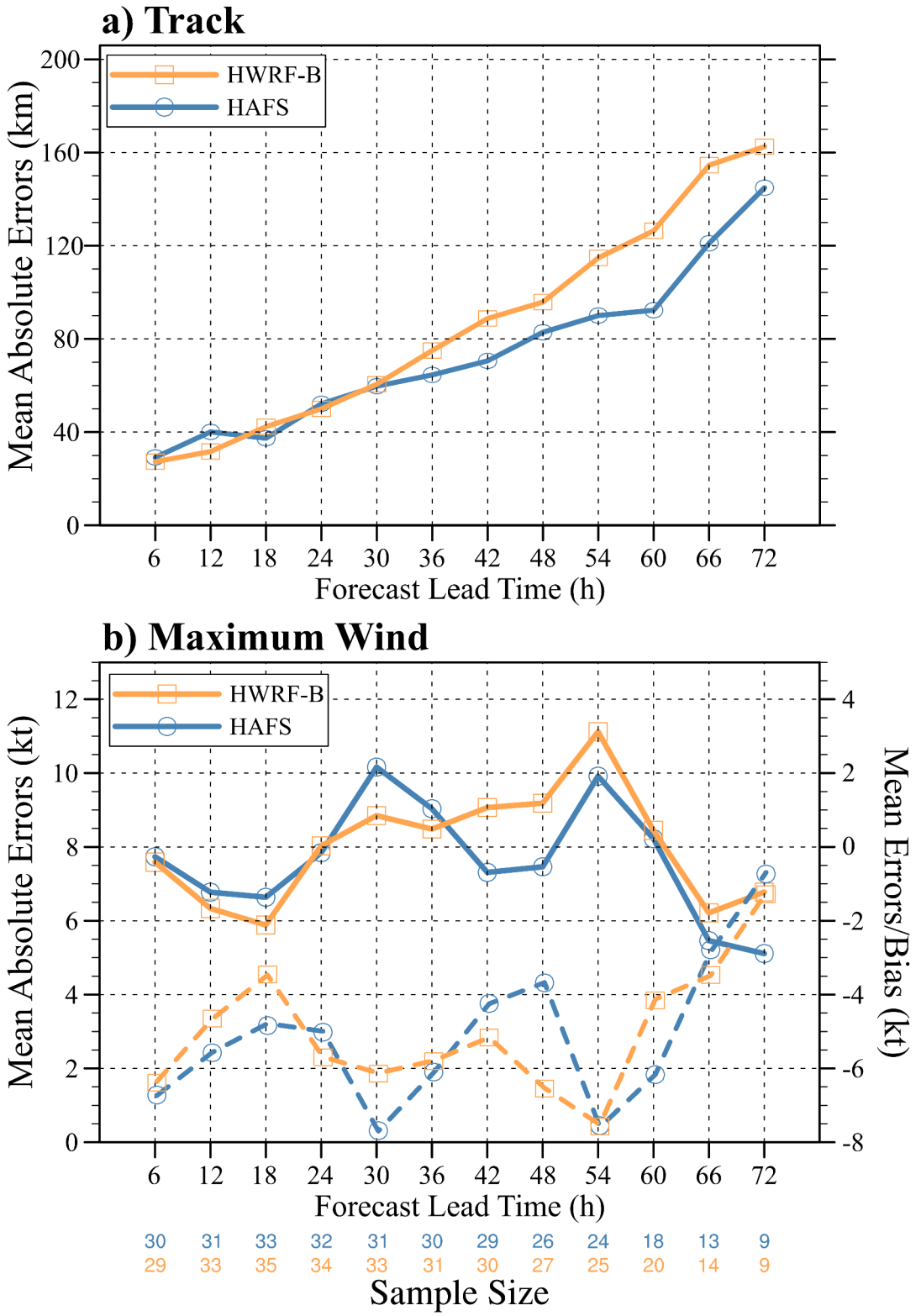

Figure 1. Verification of 6-72-h HWRF-B (orange square) and HAFS (blue circle) forecast errors for landfalling TCs included in this study. Specifically: a) mean absolute track errors in units of nautical miles, and b) mean absolute intensity (i.e. maximum wind) errors (solid) and mean intensity errors/bias (dashed) in units of knots. Values below panel (b) indicate the sample sizes for HAFS (top row) and HWRF-B (bottom row) at each forecast lead time are shown beneath the panels. Errors are computed relative to the HURDAT2.

3. Methods

a. Object-based verification method

To perform an object-based analysis on the precipitation fields for the forecast and the observed data, it is preferable to use a common grid due to minor sensitivities of the spatial metrics to the resolution. Here, the HAFS, HWRF-B, and Stage IV data are interpolated onto a grid with 5-km resolution using the nearest neighbor method. Currently, there are no studies that investigate how different interpolation methods affect object-based analyses, but Accadia et al. (2003) found the nearest neighbor method to be more suitable than the bilinear method when performing high-resolution grid transformations. In addition to the regridding, a mask is applied to conceal forecast and observed data outside of 150 km from the coastal NEXRAD sites towards the surrounding bodies of water due to the limitations of radar range and the lack of rain gauge data over water. This mask is generated using a 2-minute resolution topography map to identify ocean points (where elevation equals 0 meters) within 150 km of radar sites; ocean points farther than 150 km from a radar site are removed or masked. That mask was then interpolated to the 5-km common grid.

The precipitation objects are obtained by first setting a search radius and then thresholding the precipitation values. Precipitation data are only evaluated within 600 km of the TC center. Many studies use 500 km as a search radius for TCs over ocean (Lonfat et al. 2004; Hernández Ayala and Matyas 2016), but since this study analyzes landfalling TCs, the search radius is increased to account for the expansion of the rainbands that occurs during landfall and extratropical transition, similar to other studies of TC landfall (Zick and Matyas 2016; Matyas et al. 2018; Kirkland and Zick 2019). For the TC center, we use 3-hourly spline interpolated HURDAT2 center positions for the observed precipitation, and we use the model forecast storm center for the forecast precipitation.

A binary image is then established by applying 2, 5, and 10 mm hr^{-1} rain rate thresholds to the precipitation field (e.g., Figure 2). Precipitation intensities that are greater than or equal to the threshold are assigned a value of one while the intensities that are less than the threshold

are assigned a value of zero. The rain rate of 2 mm hr^{-1} approximately represents the threshold between precipitation resulting from warm clouds and cold clouds (Lau and Wu 2003); therefore, the 2 mm hr^{-1} threshold is chosen to evaluate the model's ability to capture cold rain processes that are stratiform (non-convective) in nature. While this region contains both stratiform and convective precipitation, it is composed of primarily stratiform precipitation with smaller embedded regions of convective precipitation. The 5 and 10 mm hr^{-1} rain rates represent two approximate thresholds between stratiform and convective precipitation. Even though deep convection in TCs can cause convective rain rates to occur below the 10 mm hr^{-1} threshold (Tokay et al. 1999), this research focuses on landfalling TCs, so the 10 mm hr^{-1} threshold more accurately captures convective precipitation over land (Schumacher and Houze 2003). Assessing stratiform and convective precipitation provides an opportunity to evaluate any differences in the model's capabilities when forecasting low versus high rain rates.

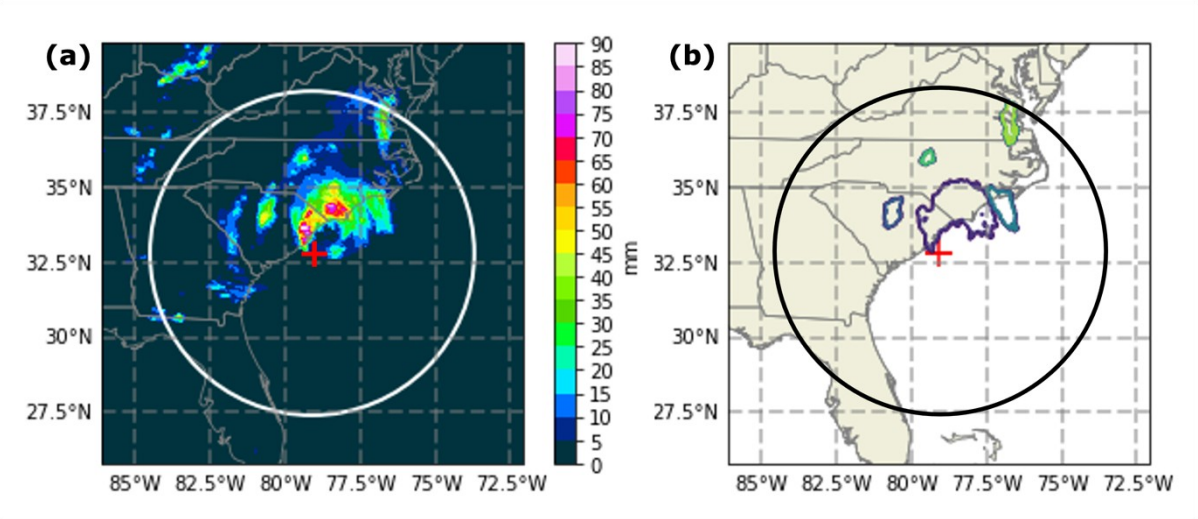

Figure 2. Illustration of methodology to obtain precipitation objects using (a) the Stage IV precipitation field for Hurricane Isaias accumulated over 3 hours and valid 03 UTC 4 August 2020 and (b) the precipitation objects derived from (a) using a search radius of 600 km and a precipitation threshold of 5 mm hr^{-1}. The white and black circles represent the 600-km search radius, and the red cross indicates the TC center.

After the thresholding process, the precipitation objects are defined by the connected pixels (those that share sides) and have a value of one. Lastly, minimum areas of 10,000 km^2 for the 2 mm hr^{-1} threshold objects and 1,000 km^2 for the 5 and 10 mm hr^{-1} threshold objects are applied to focus on mesoscale precipitation regions. Objects tend to be smaller at the higher rain rate thresholds, so a stricter minimum area threshold is implemented. Pinto et al.

(2015) perform a similar procedure for isolating mesoscale convective systems. Figure 2 shows a visual representation of the object identification process described above. The smaller objects in the field and precipitation outside of the search radius are eliminated, leaving only precipitation objects that meet the specified rain rate and area thresholds.

A variety of shape metrics, based on the geometric attributes of the forecast and observed objects, are utilized in this study. Table 3 displays the computed shape metrics, described here, for the precipitation field in Figure 2. All shape metrics are unitless unless otherwise specified. The shape metrics fall under two broad categories which include general metrics and location metrics. The *general metrics* (Table 4) are familiar measures that have formulas composed of geometric attributes (see equations in Figure 3); many have been used in previous studies employing object-based methods to measure TC features (Matyas 2007; Zick and Matyas 2016; Matyas et al. 2018; Zick 2020; Zick et al. 2022). Elongation is a measure of the ratio of the minor to the major axis, and it gives insight into whether the precipitation objects are more circular (near zero) or oval (near one). Solidity is a measure of how much area a precipitation object fills with respect to its convex hull, the smallest convex shape fitted to the object. The higher values represent precipitation fields with fuller objects while the lower values represent fields with emptier objects. Closure measures the distribution of the precipitation about the TC center. To achieve a high value, precipitation must enclose or nearly enclose the TC center. Closure can also be modified to limit the radial extent to smaller circles (or distances from the TC center). In this study, we use 100-, 150-, and 500-km variations of closure. The two lower search radii exclude the precipitation from distant rainbands, which enables evaluation of the closure of the TC inner core. The dispersion metric measures the precipitation object distances from the TC center, with larger objects weighted more heavily in the calculation. Higher values show that the objects within a precipitation field are farther from the TC center while lower values show that the objects are closer to the TC center. Fragmentation is a measure of the patchiness of the TC precipitation, which results from breaking up the precipitation field (i.e., increasing number of objects) and patchiness in the precipitation objects themselves (i.e., the emptiness of the objects). Higher values of this metric show that the TC precipitation is more fragmented while lower values show cohesiveness of the TC precipitation. Supplementary metrics that do not depend on an equation include the Sum of Areas and Sum of Perimeters, which involve summing their respective values over all objects in the field.

Metric	**Value**	**Metric**	**Value**	**Metric**	**Value**
Elongation	0.35	Solidity	0.75	Closure (100 km)	0.16
Closure (150 km)	0.31	Closure (500 km)	0.38	Dispersion	0.36
Fragmentation	0.61	Sum of Areas	2043 pixels	Sum of Perimeters	534.77 pixel sides
Weighted Centroid Longitude	-78.3°	Weighted Centroid Latitude	34.5°	Max Precipitation Bearing	347.9°
X-Displacement	71.5 km	Y-Displacement	188.8 km		

344 Table 3. General and location shape metrics corresponding to the objects shown in Figure
345 2.
346

Metric	Description
Elongation	Calculates the average elongation of the objects within a precipitation field
Solidity	Computes the fraction of the sum of all object areas to the sum of their convex hull areas
Closure	Determines the distribution of precipitation around the TC center
Dispersion	Measures the spread of the objects from the TC center
Fragmentation	Provides a measure of how fragmented the objects are themselves and as a whole
Sum of Areas	The number of pixels that form an object
Sum of Perimeters	The number of pixels that form the boundary of an object

Table 4. Descriptions of the "General" metrics.

Metric	Near 0	Near 1	Equation
Elongation	Circular 	Elliptical 	$\sum_{i=1}^{N} \frac{Area_i}{Total\ Area} * \frac{Major\ axis\ length_i - Minor\ axis\ length_i}{Major\ axis\ length_i}$
Solidity	Empty 	Filled 	$\frac{Total\ Area}{Total\ Convex\ Area}$
Closure	Exposed 	Enclosed 	$\frac{\#\ of\ 1^\circ\ slices\ with\ precipitation}{360}$
Dispersion	Central 	Dispersed 	$\sum_{i=1}^{N} \frac{Area_i}{Total\ Area} * \frac{Distance\ from\ TC\ center_i}{Search\ Radius}$
Fragmentation	Cohesive 	Fragmented 	$1 - (Connectivity * Solidity)$

Figure 3. Interpretations of the General metrics with corresponding equations. N represents the total number of objects in a precipitation field and the total area/convex area is the sum of the areas/convex areas of all objects in a precipitation field.

There are numerous *location metrics* related to the location of the objects/pixels in a precipitation field. The metrics in this category are the displacement in the x-and y-directions and the maximum precipitation bearing. The displacement metrics depend on the centroid, which is the center of mass of an object, represented in longitude and latitude values. When multiple objects exist, we use a weighted centroid, which is the average centroid position weighted by the area of the objects in the binary precipitation field. The x- and y-displacement metrics are east-west and north-south components of the displacement, respectively. These displacement metrics help determine whether the model has a bias associated with the placement of TC precipitation and may also give insight into along-track and across-track errors with the simulated TC track. Finally, to gain insight into any significant differences between where the model forecasts the TC maximum precipitation and where the observed data places this maximum, the bearing from the TC center to the maximum precipitation pixel is evaluated. When more than one maximum precipitation pixel is identified, we use the average position of those pixels.

After the shape metrics are calculated, these metrics are compared between the forecast and observed precipitation fields for each storm using the Mann Whitney *U* test. According to Wilcoxon (1992), the Mann Whitney *U* test is a nonparametric statistical test that compares the distributions of two independent groups, which are the model and observed shape metrics for this analysis, based on the mean ranks of the data. If the distributions of the two independent groups are assumed to be the same, then the test compares differences in the medians instead of the distributions. After the test is performed, the null hypothesis is rejected for shape metrics that are associated with *p*-values that are equivalent to or less than a significance level of 0.05 (also referred to as "significant metrics" in this study), which indicates disagreement between the model and the observed shape metric values. When differences are found in a particular shape metric, we perform a post-hoc comparison to evaluate whether the model is over- or under-forecasting that metric. Due to the mask, some of the TC precipitation may be removed, affecting the shape of the precipitation. However, as long as the TCs are located in similar locations, it will affect the model and observed precipitation similarly. Here, we focus on model verification for the 3-72 hour forecasts, when position differences are generally small (see Figure 1). In this study, storm locations are similar and the mask does not affect the results, but future studies should keep this in mind when a mask is needed due to data coverage.

The model objects are also assessed based on the lead time of the forecasts. The shape metrics are compared between the forecast and observed data for each forecast hour. The maximum forecast hour is 72 hours; forecasts beyond three days are not included in this analysis. The Mann Whitney *U* test is used, as with the general object assessment, to identify significant differences between the model and observed shape metrics at each lead time.

b. Point-based verification

Point-based verification methods are included in this study to determine how these approaches perform compared to the object-based verification methods. Section 5 shows that point-based methods are less representative of the model performance, so these metrics are not used in the full object-based analysis of the HAFS and HWRF-B precipitation forecasts in section 6. The point-based metrics include Spearman's rank correlation, categorical statistics, and error metrics.

The Spearman's rank correlation (Spearman 1961) is a nonparametric test used to quantify the strength of the monotonic relationship between two datasets. The values of the

Spearman's "rho" coefficients fall in the range [-1, 1] in which negative one represents a perfect negative relationship, positive one represents a perfect positive relationship, and zero represents no relationship. Categorical statistics are derived from counts of hits, misses, false alarms and correct negatives, which make up the contingency table (Wilks 2011). In this study, we use probability of detection (POD), false alarm ratio (FAR), and equitable threat score (ETS). The POD represents the fraction of observed precipitation that was correctly forecasted (Wilks 2011). The FAR is the fraction of the forecasted precipitation that was not observed. ETS is a variation of the threat score (also known as the critical success index), which measures the fraction of forecasted events that were correctly predicted. The threat score is generally recommended for rare events (such as precipitation) because it measures the accuracy of a forecast after removing the influence of correct negatives (Schaefer 1990; Hamill and Juras 2006; Wilks 2011). The ETS adjusts for correctly forecasted events due to random chance (Wilks 2011). These metrics range from zero to one and are calculated for each rain rate threshold (2, 5, and 10 mm hr^{-1}).

Error metrics (Table 5) are implemented to quantify the error in the model forecasts following Wilks (2011). These metrics are calculated by comparing the forecast and observed precipitation values using multiple equations. Mean error values less than zero indicate that the forecast precipitation is under-predicted while values greater than zero indicate the forecast precipitation is over-predicted compared with the observations. The mean absolute error is the magnitude of the average forecast error and can range from zero to infinity.

Metric	Equation
Equitable threat score	$\frac{hits - hit\,s_{random}}{hits + misses + false\,alarms - hit\,s_{random}}$ where $hit\,s_{random} = \frac{(hits + misses)(hits + false\,alarms)}{total}$
False Alarm Ratio	$\frac{false\,alarms}{hits + false\,alarms}$
Probability of Detection	$\frac{hits}{hits + misses}$

Metric	Equation
Equitable threat score	$\frac{hits - hit\,s_{random}}{hits + misses + false\,alarms - hit\,s_{random}}$ where $hit\,s_{random} = \frac{(hits + misses)(hits + false\,alarms)}{total}$
False Alarm Ratio	$\frac{false\,alarms}{hits + false\,alarms}$
Mean Error	$\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)$
Mean Absolute Error	$\frac{1}{N}\sum_{i=1}^{N}¿F_i - O_i \vee ¿¿$

Table 5. Equations for the categorical statistics and error metrics, where F is the forecast data, O is the observed data, and N is the total number of pixels.

4. Point-based Metric Results from all 2020 Storms

Traditionally, TC precipitation verification has focused on point-based metrics (Tuleya et al. 2007; Villarini et al. 2011; Luitel et al. 2016; Villarini et al. 2022). We begin this analysis with an overview of model performance using some commonly-used point-based metrics described in section 3b. These include methods for dichotomous (yes/no) forecasts (probability of detection, false alarm ratio, and equitable threat score) and methods for continuous variables (Spearman's "rho" correlation, mean error, and mean absolute error). Figure 4 shows these results for the entire landfalling TC sample (provided in Table 1).

In general, the dichotomous forecast metrics indicate that forecast skill decreases with rain rate threshold (Figure 4a-c). More specifically, the POD decreases, the FAR increases, and the ETS decreases as the rain rate threshold increases. These trends are present in both the HAFS and HWRF-B model forecasts. Additionally, the HAFS model appears to have slightly greater skill (higher POD, lower FAR, higher ETS) (Figures 4a-c), though caution should be taken in comparing the two models since this study evaluates slightly different samples (Table 1). These results are consistent with other studies that find a higher false

alarm ratio and lower model forecast skill for the higher rain rate values (McBride and Ebert 2000; Zick 2020; Sierra-Lorenzo et al. 2022), a result that is generally attributed to mismatches in the location or timing of precipitation.

Point metrics based on continuous variables are also shown (Figure 4d-f). The Spearman's correlations indicate moderate to good agreement between the model and observed precipitation, with correlations for individual forecasts generally spanning 0.4–0.9 (Figure 4d). Here, higher correlations indicate that the forecast and observations are producing similar precipitation values at a given pixel. In particular, the HAFS forecast has moderate to strong correlations with the Stage IV observations, with most values falling in the 0.7-0.8 bin, indicating that the HAFS model produced precipitation patterns that were similar to the observations. Again, the HAFS forecast appears to have greater skill compared with HWRF-B, but some of that variation could be related to the slightly different samples for the two models. These moderate to high correlations are encouraging since the specific location of heavier precipitation is important for flood forecasting. Mean error (Figure 4e) can be used to understand the model bias compared with observations. The HWRF-B forecasts have a slight negative bias with most of the forecasts producing precipitation that is about 0-1 mm hr^{-1} lower than the observations. The HAFS forecasts exhibit a more evenly distributed mean error around the center, which indicates a bias closer to zero for its sample. Mean absolute errors for the two models are similar and predominantly fall between 1-3 mm hr^{-1} (Figure 4f).

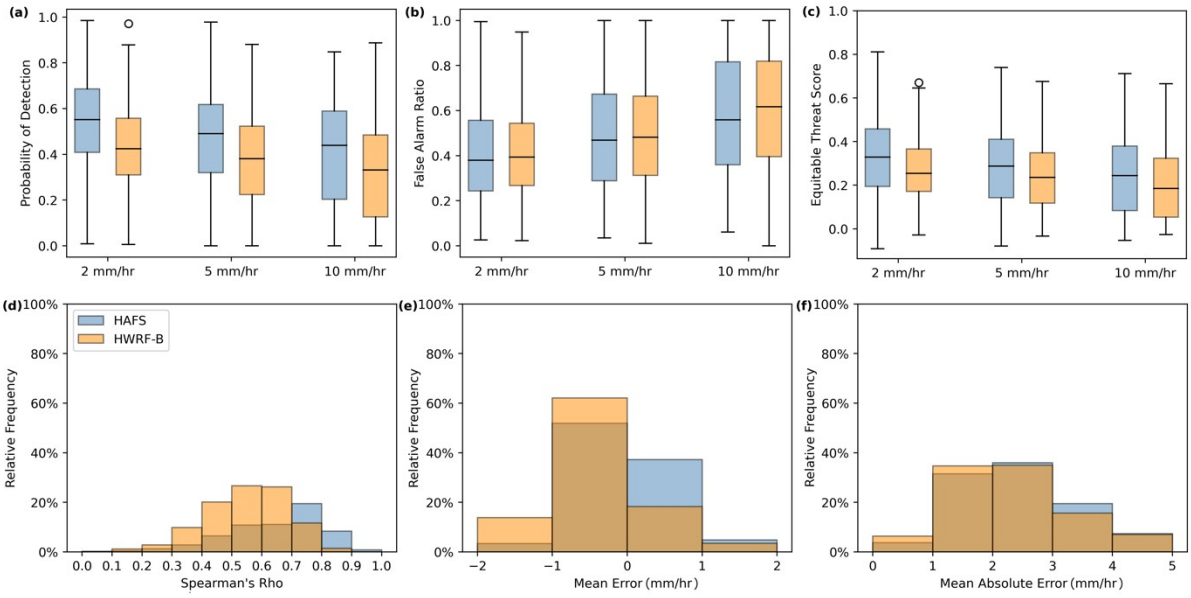

Figure 4: (a-c) Box plots and (d-f) histograms showing point-based verification metrics for all 2020 landfalling TCs for the 3-72 h precipitation forecasts from the the HAFS and HWRF-B models compared with Stage IV observations.

5. Results from Case Studies of Hurricane Isaias and Hurricane Laura

In this section, we present two case studies to illustrate the model evaluation of the precipitation forecasts of Hurricane Isaias and Hurricane Laura. Here, the object-based and point-based verification methods are demonstrated and compared. Specifically, we use the results in this section to show that point-based metrics can be misleading when there are location errors. For Hurricane Laura, the point-based metrics indicated lower performance compared to Hurricane Isaias. Yet, the object-based metrics indicate better overall performance in Hurricane Laura, which means that the Hurricane Laura precipitation structure is similar between the model and observed fields despite the location errors indicated by the point-based metrics. This section also illustrates how the object-based metrics can be used to evaluate model forecast precipitation in the two case studies.

a. Hurricane Isaias Forecast Evaluation

As summarized by Latto et al. (2021), Hurricane Isaias made landfall near Ocean Isle Beach, North Carolina at category 1 intensity on 3 August 2020. The hurricane quickly weakened into a tropical storm and traveled northeastward, affecting states farther up the East Coast. On 5 August, the storm became an extratropical system. The highest precipitation totals from this system occurred in parts of Virginia, Maryland and Pennsylvania (Roth 2021). The period of study for Hurricane Isaias used in this analysis begins 2100 UTC 3 August, near landfall, and ends 0000 UTC 5 August, when the storm became extratropical. A total of 88 and 79 (3-hourly) model output times were used from HAFS and HWRF-B forecasts of Hurricane Isaias, respectively (Table 1).

1) POINT-BASED VERIFICATION

The point-based metrics indicate overall similar skill for the HAFS and HWRF-B precipitation forecasts for Hurricane Isaias (Figure 5), with a note that the samples for the two models are not homogeneous and therefore should be compared with caution. The POD values suggest that both models are forecasting more hits than misses (Figure 5a), and the FAR values are relatively low at the lower thresholds and relatively high at the higher thresholds (Figure 5b). The trends in FAR and POD values as the threshold increases indicate that the models performed more poorly at greater thresholds. The ETS values also indicate moderately decreasing model skill with rain rate threshold (Figure 5c). The correlation values

show that the forecast and observed fields were well correlated, but these values are slightly higher for the HAFS model than the HWRF-B model (Figure 5d). The mean error values show that both models slightly underestimated the precipitation of Hurricane Isaias (Figure 5e), but biases are slightly improved compared with the larger TC sample (Figure 4e) . The mean absolute error shows similar error between the HAFS and HWRF-B models with many values falling between 1 and 4 mm (Figure 5f), which is higher than the mean absolute errors for the entire landfalling TC sample (Figure 4f). However, correlations are generally higher for Hurricane Isaias compared with the larger TC sample (Figures 4d, 5d), suggesting better pattern matching in Hurricane Isaias compared with the other TCs in this study.

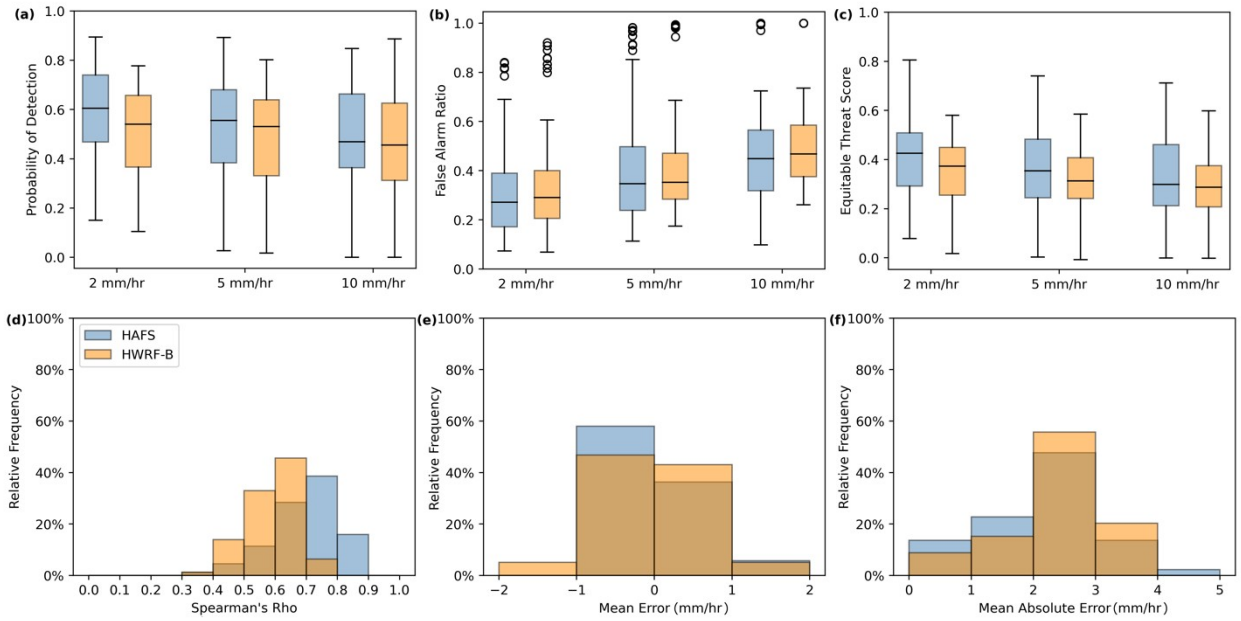

Figure 5: (a-c) Box plots and (d-f) histograms showing point-based verification metrics for Hurricane Isaias (2020) for the 3-72 h precipitation forecasts from the HAFS and HWRF-B models compared with Stage IV observations.

2) Object-Based Verification

For Hurricane Isaias, there are many more significant metrics identified by the Mann Whitney *U* test for the HAFS forecasts compared with the HWRF-B forecasts (Table 6). A larger number of significant metrics indicates more discrepancies between the shape of the forecast and observed fields and, thus, a less accurate forecast. The number of significant metrics increases as the rain rate threshold increases for the HAFS model, while there is no similar trend for the HWRF-B model (Table 6). At the 2 mm hr^{-1} threshold, both models significantly under-forecasted the sum of areas, dispersion, and x-and y-displacement metrics (Table 6, Figure 6a, Figure 7d-f). Additionally, the HWRF-B model overestimates the 150-

km closure (Figure 6a, Figure 7d-e). At the 5 mm hr^{-1} threshold, the HAFS model performed similarly to the 2 mm hr^{-1} threshold in addition to overestimating 100- and 500-km closure (Table 6, Figure 6a, Figure 7g,i). In contrast, the HWRF-B model performed the best at the 5 mm hr^{-1} threshold by only under-forecasting the dispersion (Table 6, Figure 6b, Figure 7h-i). At the 10 mm hr^{-1} threshold, the HAFS model still struggled with capturing the dispersion, x-displacement, y-displacement, and closure metrics, but the fragmentation was also under-forecasted, and the solidity was over-forecasted (Table 6, Figure 6c, Figure 7j,l). Like the HAFS model, the HWRF-B model had issues with underestimating the fragmentation and x-displacement and overestimating the 150-km and 500-km closure at the 10 mm hr^{-1} threshold (Table 6, Figure 6c, Figure 7k-l).

In general, across all thresholds, there is a tendency for both models to forecast precipitation that is too central/compact and too enclosed around the TC center compared with the observations. Typically, higher compactness and closure would be associated with a more intense TC in the model compared with the observations. However, for Isaias, both models have a negative intensity bias (Table 2). This is an interesting result that will be examined further in section 6. Additionally, the x- and y-displacement metrics suggest location biases in the precipitation forecast. These location biases might be associated with slightly higher mean absolute track errors in Hurricane Isaias compared with other 2020 landfalling TCs (Table 2). Positional differences will also be explored further in section 6.

	HAFS			HWRF-B		
Metric	2 mm hr^{-1}	5 mm hr^{-1}	10 mm hr^{-1}	2 mm hr^{-1}	5 mm hr^{-1}	10 mm hr^{-1}
Solidity	0.93	0.14	**<0.01**	0.95	0.67	0.08
Elongation	0.97	0.92	0.3	0.19	0.26	0.83
Closure (100 km)	0.11	**0.01**	**<0.01**	0.53	0.4	0.13
Closure (150 km)	**0.03**	**0.01**	**<0.01**	0.68	0.18	**0.02**
Closure	0.06	**0.03**	**0.01**	0.65	0.1	**0.01**

(500 km)						
Dispersion	**0.03**	**0.01**	**<0.01**	**0.01**	**0.04**	0.09
Fragmentation	0.55	0.57	**<0.01**	0.42	0.87	**0.03**
Sum of Areas	**0.02**	0.16	0.12	**0.01**	0.54	0.15
Sum of Perimeters	0.16	**0.05**	0.08	0.06	0.18	0.93
X-Displacement	**0.02**	**<0.01**	**<0.01**	**0.03**	0.14	**0.04**
Y-Displacement	**0.04**	**0.01**	**0.01**	**0.03**	0.15	0.09

533 Table 6. Mann Whitney *U* test results (*p*-values) from comparisons between model
534 (HAFS and HWRF) and Stage IV metrics calculated for Hurricane Isaias (2020). Significant
535 *p*-values (<= 0.05) are indicated with a bold-faced font style.

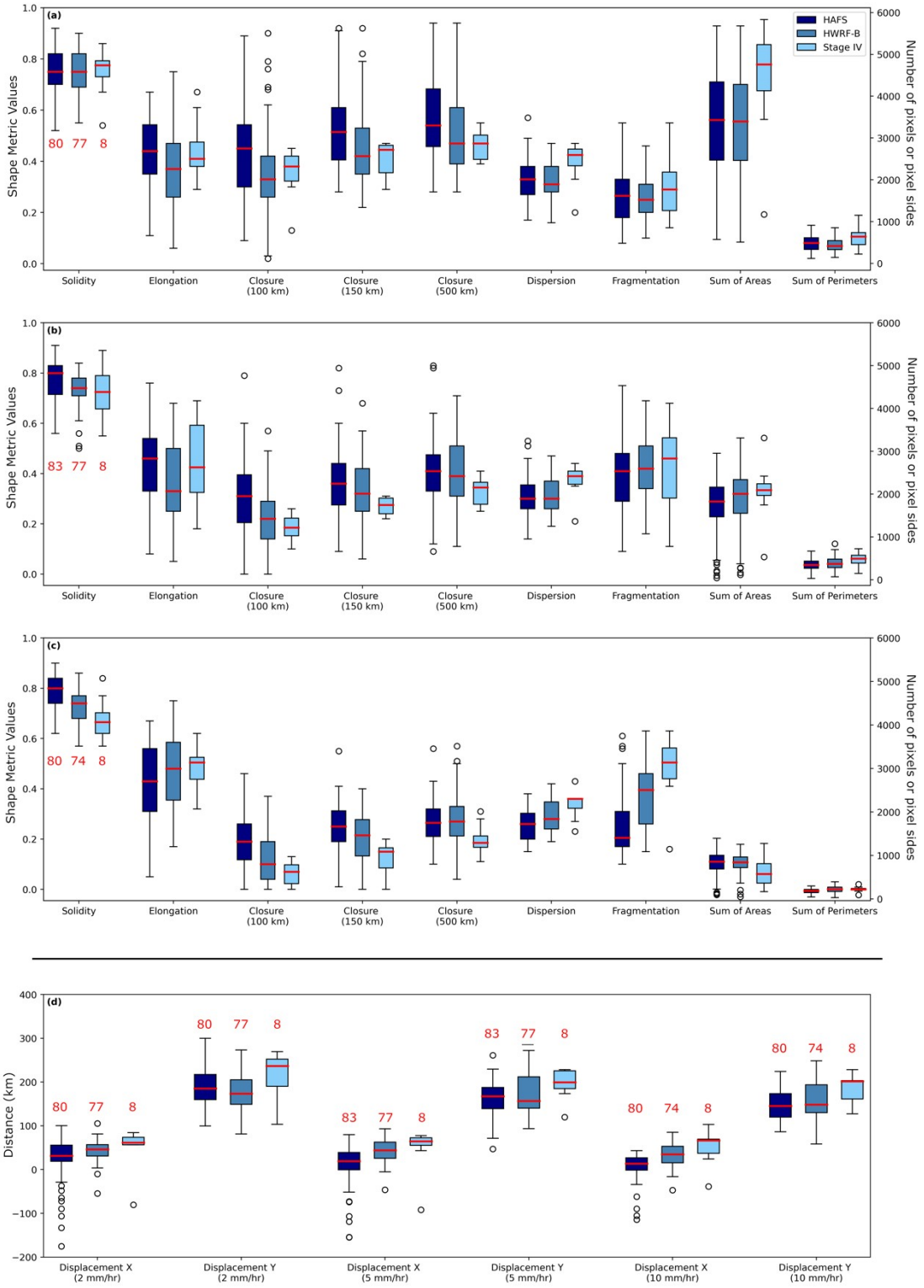

Figure 6: Box plots showing object-based metrics for the HAFS and HWRF-B models compared with Stage IV for Hurricane Isaias (2020) for general metrics at (a) 2 mm hr^{-1}, (b) 5 mm hr^{-1}, and (c) 10 mm hr^{-1} thresholds and for (d) location metrics at all three thresholds. In (a) – (c), shape metric values range from 0 to 1 and the values are indicated on the left axis, while sum of areas and sum of perimeters are the number of pixels and values are indicated on the right axis. In (d), shape metric values are distances as indicated on the left axis.

Sample sizes for (a) – (c) are equal across all metrics and labeled for the left-most solidity metric only. Sample sizes for (d) are labeled for all metrics.

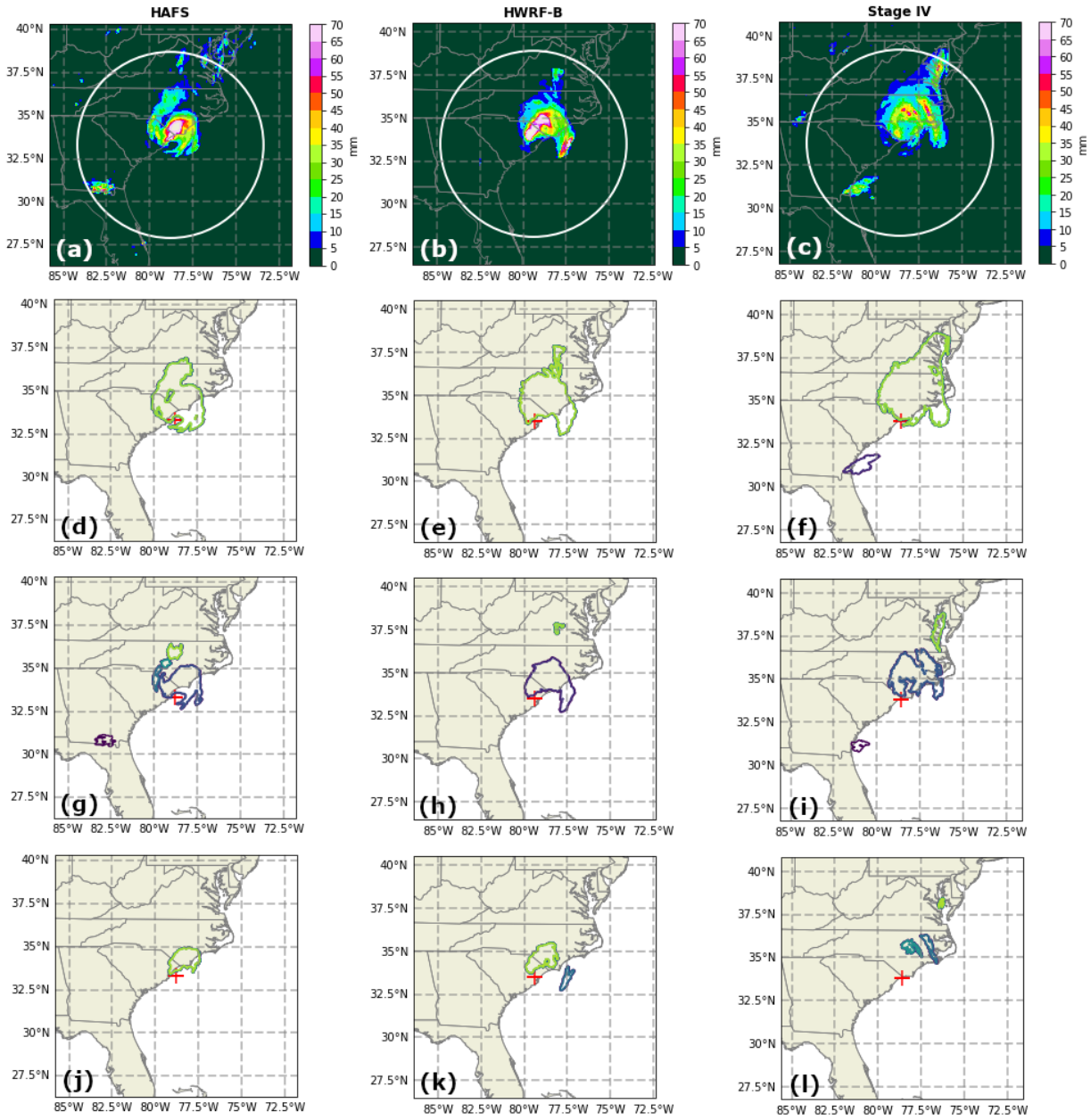

Figure 7: Precipitation and precipitation objects for the Hurricane Isaias forecast initialized 00 UTC 2 Aug 2020 and valid 06 UTC 4 Aug 2020. (a-c) show full 3-hourly precipitation, (d-f) show corresponding 2 mm hr^{-1} objects, (g-i) show corresponding 5 mm hr^{-1} objects, and (j-l) shows corresponding 10 mm hr^{-1} objects for (left) HAFS, (middle) HWRF-B and (right) Stage IV. The white circle represents the search radius, and the red cross indicates the TC center.

b. Hurricane Laura Forecast Evaluation

Pasch et al. (2021) report that Hurricane Laura made landfall near Cameron, Louisiana as a category 4 storm on 27 August 2020 and then weakened into a tropical storm over Arkansas

on 28 August as it traveled inland. The storm acquired a northeastward track over northern Arkansas and became a remnant low over Kentucky on 29 August. The greatest rainfall accumulations occurred in Louisiana and portions of Arkansas. The study period for Hurricane Laura spans from 0600 UTC 27 August, at landfall, to 0600 UTC 29 August, when the storm transitioned into a remnant low. A total of 158 and 162 (3-hourly) timesteps were used in the evaluation of the HAFS and HWRF-B model forecast of Hurricane Laura, respectively (Table 1). It is important to note that the Lake Charles and Fort Polk radars were down during the entire study period, which may contribute to degraded forecast performance due to an underestimation of Stage IV rainfall in central and southeastern Louisiana. An underestimation of Stage IV rainfall could lead to underestimation of areal coverage for our object-based metrics, as well as reduced performance in the other spatial metrics (e.g., lower closure if eyewall is not sampled).

1) Point-Based Verification

For Hurricane Laura, the HAFS model overall performed better than the HWRF-B model in forecasting precipitation as shown by the point-based metrics (Figure 8). In this case, the samples are very similar (Table 1), but the two models should still be compared with caution. The POD indicates better performance in forecasting hits for the HAFS model (Figure 8a). However, the FAR values indicate that there were slightly more false alarms forecasted for the HAFS model than the HWRF-B model (Figure 8b). The ETS value distributions are generally similar for the two models (Figure 8c). As with the Isaias forecast, the trends in these three point-based metrics show a decrease in model skill with higher rain rate thresholds (Figure 8a-c), consistent with more mismatches in location for the higher rain rate values. The correlations are overall moderate for both models, but slightly higher for the HAFS model (Figure 8d). A tendency for the models to under-forecast precipitation is shown by the mean error values (Figure 8e), but this underestimation is more evident for the HWRF-B model. These mean errors indicate a stronger negative bias in the Hurricane Laura forecasts compared with the entire landfalling TC sample (Figures 4e, 8e). The mean absolute error values are similar for both models with most values ranging between 1 and 3 mm, which is similar to the entire TC sample (Figures 4f, 8f). Compared with the Isaias forecast, correlations between the model forecast and the observations are much lower in Hurricane Laura (Figures 5d, 8d). Additionally, the correlations for Hurricane Laura are slightly lower than the larger TC sample (Figures 4d, 5d), suggesting a poorer pattern match compared with the observations.

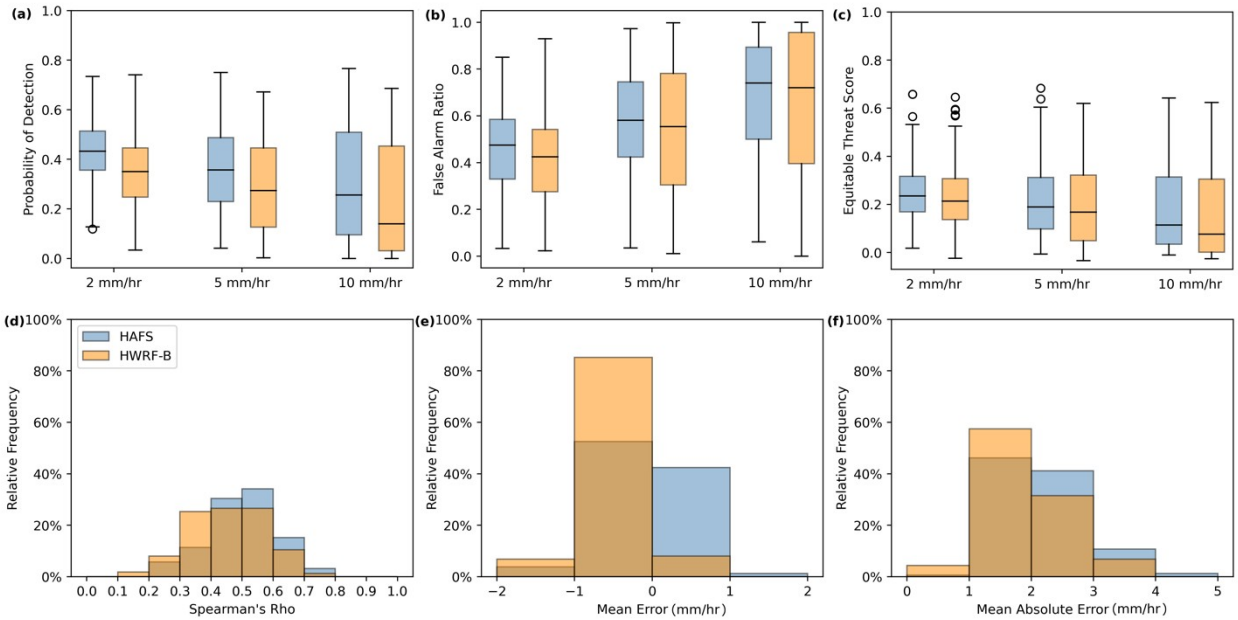

Figure 8: (a-c) Box plots and (d-f) histograms showing point-based verification metrics for Hurricane Laura (2020) for the 3-72 h precipitation forecasts from the HAFS and HWRF-B models compared with Stage IV observations.

2) Object-Based Verification

There are more significant metrics resulting from the Mann Whitney *U* test for the HWRF-B forecast than the HAFS forecast for Hurricane Laura (Table 7), indicating that the HAFS forecast precipitation structure was more like the observations. The HAFS and HWRF-B samples are very similar for Hurricane Laura (Table 1), but model comparisons should still be treated with caution. For both models, the number of significant metrics decreases as the threshold increases, indicating that accuracy increases with higher rain rates (Table 7). At the 2 mm hr^{-1} threshold, both models under-forecasted the sum of areas, sum of perimeters, and dispersion (Table 7, Figure 9a, Figure 10d-f). Additionally, the HWRF-B model underestimated the fragmentation and overestimated the y-displacement (Table 7, Figure 9a, Figure 10e-f). The only significant metric at the 5 mm hr^{-1} threshold for the HAFS model is the solidity, which was over-forecasted (Table 7, Figure 9b, Figure 10g,i). Similar to the 2 mm hr^{-1} threshold, the HWRF-B model underestimated the dispersion and fragmentation and overestimated the y-displacement at the 5 mm hr^{-1} threshold (Table 7, Figure 9b, Figure 10h-i). At the 10 mm hr^{-1} threshold, there are no significant metrics for the HAFS model (Table 7), but the HWRF-B over-forecasted the y-displacement as seen with the lower thresholds (Table 7, Figure 9c, Figure 10k-l). The lack of radar coverage is possibly affecting the Stage IV observations during part of the study period. Specifically, a rainband extending to the southeast of the TC center might be underestimated in the Stage IV product

611 (Figure 10c), which would also impact the solidity and fragmentation calculations. Still, the
612 object-based metrics indicate good agreement between the model forecast and observed
613 precipitation, especially for the HAFS model (Table 7). Visual inspection of Stage IV data,
614 along with comparisons to satellite-derived rain rates, indicates that there is likely a minimal
615 impact of the radars going down in this case, except for during a short 3-hour window around
616 0600 – 0900 UTC 27 August. After that time, a dry slot forms within the Fort Polk and Lake
617 Charles radar coverage regions, which likely explains the minor impact on our results.

618 For the Hurricane Laura forecast, the object-based metrics generally indicated a good
619 match between the TC structure in the model and the observations, which is consistent with
620 relatively low intensity errors in the model forecasts for this storm (Table 2). More structural
621 differences are present at the lower rain rate thresholds for both models, which suggest some
622 issues with the representation of stratiform rain. There is also a consistent underestimation of
623 y-displacement by HWRF-B, indicating that this model predicted precipitation that was too
624 far south compared with the observations. Track errors are similar in the HAFS and HWRF-B
625 forecasts of Hurricane Laura (Table 2), so these locational differences might be related to a
626 southward bias in the position of rainbands with respect to the TC center, which is likely
627 contributing to the higher dispersion values in the HWRF-B forecast as well (Figure 9a-c).

	HAFS			HWRF-B		
Metric	2 mm hr^{-1}	5 mm hr^{-1}	10 mm hr^{-1}	2 mm hr^{-1}	5 mm hr^{-1}	10 mm hr^{-1}
Solidity	1	**0.05**	0.63	0.44	0.47	0.38
Elongation	0.52	0.68	0.65	0.32	0.63	0.73
Closure (100 km)	0.41	0.56	0.6	0.69	0.75	0.63
Closure (150 km)	0.63	0.88	0.68	0.92	0.99	0.69
Closure (500 km)	0.65	0.53	0.31	0.27	0.63	0.23

Dispersion	**0.03**	0.19	0.3	**<0.01**	**0.01**	0.09
Fragmentation	0.45	0.82	0.31	**0.01**	**<0.01**	0.75
Sum of Areas	0.05	0.97	0.65	**<0.01**	0.27	0.94
Sum of Perimeters	**0.03**	0.58	0.56	**<0.01**	**<0.01**	0.98
X-Displacement	0.77	0.26	0.24	0.37	0.8	0.54
Y-Displacement	0.28	0.2	0.11	**0.05**	**0.02**	**0.04**

628 Table 7. Mann Whitney *U* test results (*p*-values) from comparisons between model
629 (HAFS and HWRF) and Stage IV metrics calculated for Hurricane Laura (2020). Significant
630 *p*-values (<= 0.05) are indicated with a bold-faced font style.

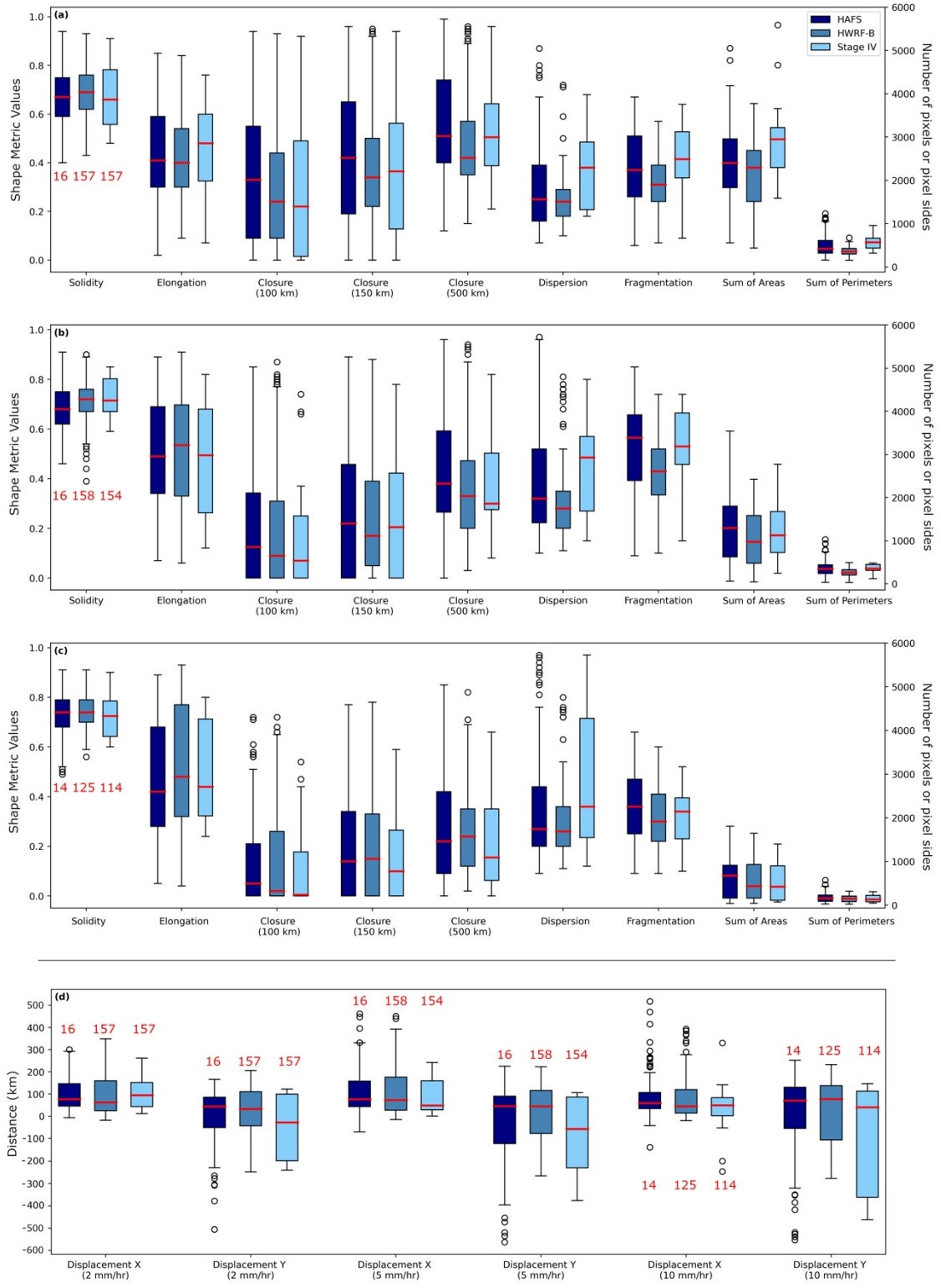

Figure 9: Box plots showing object-based metrics for the HAFS and HWRF-B models compared with Stage IV for Hurricane Laura (2020) for general metrics at (a) 2 mm hr^{-1}, (b) 5 mm hr^{-1}, and (c) 10 mm hr^{-1} thresholds and for (d) location metrics at all three thresholds. In (a) – (c), shape metric values range from 0 to 1 and the values are indicated on the left axis, while sum of areas and sum of perimeters are the number of pixels and values are indicated on the right axis. In (d), shape metric values are distances as indicated on the left axis.

Sample sizes for (a) – (c) are equal across all metrics and labeled for the left-most solidity metric only. Sample sizes for (d) are labeled for all metrics.

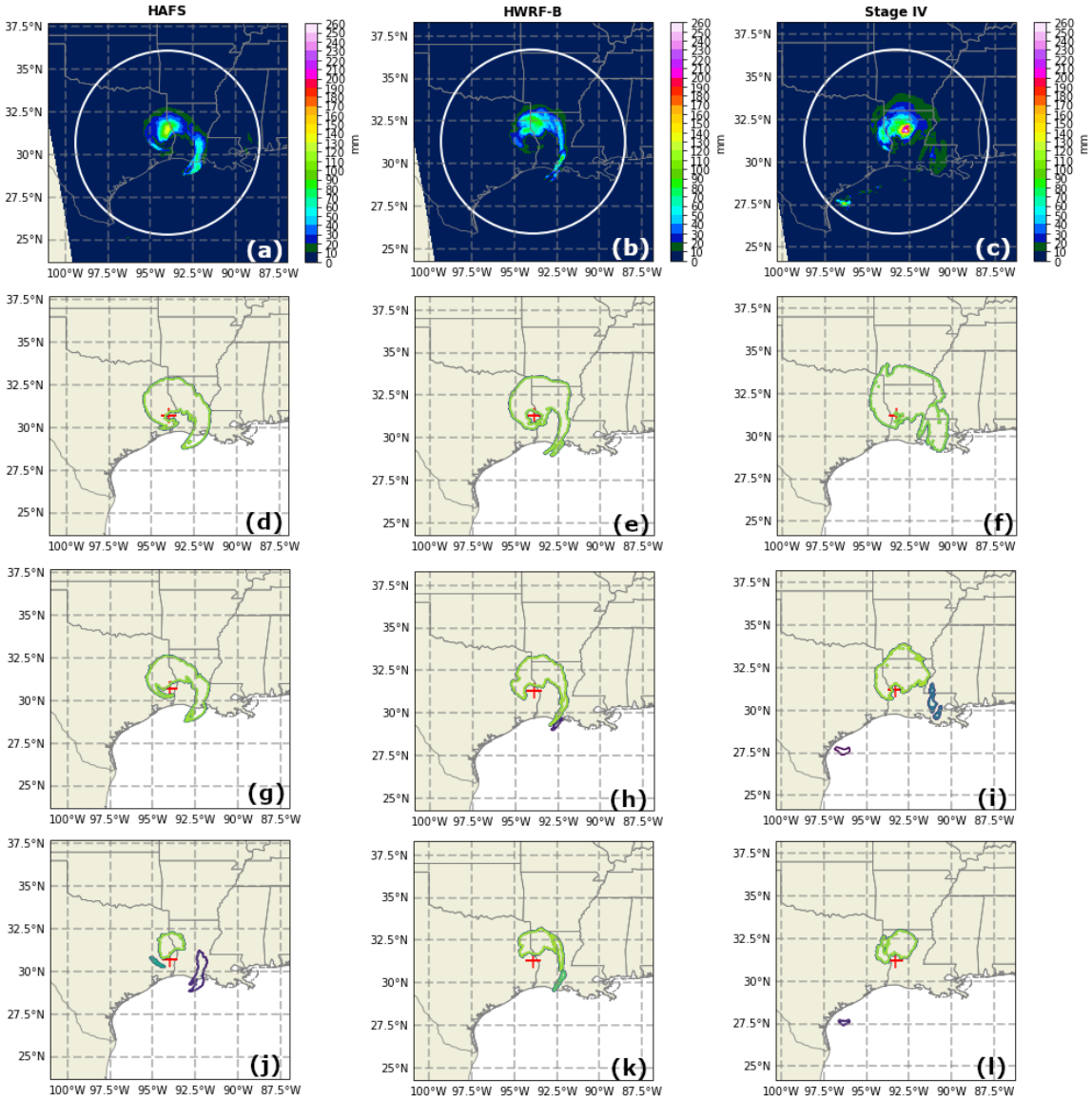

Figure 10: Precipitation and precipitation objects for the Hurricane Laura forecast initialized 12 UTC 25 Aug 2020 and valid 15 UTC 27 Aug 2020. (a-c) show full 3-hourly precipitation, (d-f) show corresponding 2 mm hr^{-1} objects, (g-i) show corresponding 5 mm hr^{-1} objects, and (j-l) shows corresponding 10 mm hr^{-1} objects for (left) HAFS, (middle) HWRF-B and (right) Stage IV. The white circle represents the search radius, and the red cross indicates the TC center.

c. Discussion

Overall, the point-based and object-based metrics communicate different results with regards to the accuracy of the models in forecasting precipitation for Hurricane Isaias compared to Hurricane Laura. The object-based metrics show better performance for the

Hurricane Laura forecast for the HAFS model and similar performance for both storms for the HWRF-B model. In contrast, most of the point-based metrics indicate that both models performed worse for the Hurricane Laura forecast, as indicated by lower correlations (Figures 5d, 8d), greater negative biases (Figures 5e, 8e), higher FAR (Figures 5b, 8b) and lower POD (Figures 5c, 8c). Interestingly, the mean absolute errors are slightly higher in Hurricane Isaias (Figures 5f, 8f), which suggests that there may be some underlying biases that are missed by the other point-based metrics. Still, the overall reduced performance in Hurricane Laura is likely due to the "double penalty" associated with point-based verification. If the forecast precipitation is displaced by even a small amount, the point-based metric scores are likely to be reduced due to misses and false alarms that are introduced by the displacement. Therefore, the models did not necessarily perform worse for the Hurricane Laura forecast than the Hurricane Isaias forecast as indicated by the point-based metrics. Instead, there tended to be more location mismatches in the Hurricane Laura forecast compared to the Hurricane Isaias forecast. The effect of the displacement errors on the point-based metrics are also shown with the reduced model skill at higher rain rate thresholds (e.g., higher FAR). The point-based metrics are likely to indicate poorer performance at greater thresholds because convective precipitation tends to be smaller in area and more isolated than stratiform precipitation, and therefore, harder to predict location-wise. The lower correlations for the Hurricane Laura forecast may also be related to missing radar data in Louisiana. Still, despite the missing radar data, the object-based forecast verification indicates better performance in the Hurricane Laura forecast. These results suggest that the object-based metrics are advantageous when there are location errors, which is a common issue in high-resolution model forecasts. The shape metrics also allow the forecasts to be evaluated in a more spatial sense than the point-based metrics. Spatial metrics gives insight into how well models perform with regards to forecasting the structure of the TC precipitation, which is useful information for model developers.

6. Object-based Results from all 2020 Storms

a. Forecast Verification in Individual Storms

This section presents the comprehensive results from implementing the object-based verification for all the 2020 storms as well as results from evaluating the effect of lead time on model performance. Figure 1a indicates that track forecast errors are initially small (< 40 km) and increase with time as expected. Absolute intensity errors consistently fall between 6

and 10 kts for both models and do not increase with forecast lead time (Figure 1b). Generally, for both models, there is a negative intensity bias (i.e., model weaker than observations) for the sample of landfalling TCs in this study. However, at longer lead times (66-72 hours), the mean errors approach zero. Note that the sample sizes decrease to fewer than ten samples at the longer lead times, and thus these data may provide a less accurate representation of the model performance (Figure 1b). Whenever possible, we compare the object-based metric results in this section with these more familiar track and intensity error metrics.

Object-based metrics indicate numerous differences between the shape and size of the precipitation field at all thresholds for both models (Table 8). Many similarities across all rain rate thresholds are also indicated. For example, the 2 mm hr^{-1} precipitation area is consistently under-forecasted by both models. The object-based analysis reveals that the two models have a similar number of shape metrics with significant differences at all rain rate thresholds. At the same time, there is greater consistency in the directional bias in the HWRF-B forecast, particularly for 2 mm hr^{-1} rain rates, which may indicate that there are more systematic issues within its forecast sample.

	Metric	Hanna	Isaias	Laura	Sally	Beta	Delta	Zeta
2 mm hr^{-1}	Sum of Areas	*,*	*A,B*	*A,B*	*A,B*	*,*	*A,B*	*,*
	Sum of Perimeters	*,*	*,*	*,B*	*,B*	*,*	*,*	*,B*
	Elongation	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Solidity	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Closure 100 km	*,*	*,*	*,*	*,*	*A,B*	*,*	**A,B**
	Closure 150 km	*,*	**A,***	*,*	*,*	*A,B*	*,*	**A,B**
	Closure 500 km	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Dispersion	*,*	*A,B*	*A,B*	*,*	*,*	*A,B*	*,B*
	Fragmentation	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	X-Displacement	*,B	*,B	*,*	*,B	*,B	*,*	*,*
	Y-Displacement	*A,**	*A,B*	*,**B**	*,*	*,*	*,*	**A,***
5 mm hr^{-1}	Sum of Areas	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Sum of Perimeters	**A,B**	*A,**	*,*	*,B*	*,*	*,B*	*,*

	Elongation	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Solidity	*A*,*	*,*	*A*,*	*A*,*	*,*	*,*	*,*
	Closure 100 km	*,*	**A**,*	*,*	*,*	*A*,*	*,*	*,*
	Closure 150 km	*,*	**A**,*	*,*	*,*	*A*,*	*,*	*,*
	Closure 500 km	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Dispersion	*,*	*,*B*	*,*B*	*,*	*,*	*,*B*	*,*B*
	Fragmentation	*,*	*,*	*,*	**A**,*	**A**,*	*,*	*,*
	X-Displacement	*,*B*	*A*,*	*,*	*,*B*	**A**,*B*	*,*	*,*
	Y-Displacement	*A*,*	*A*,*	*,*	*,*	*,*	*,*	*,*
10 mm hr^{-1}	Sum of Areas	*,*	*,*	*,*	*,*B*	*,*B*	*,*	*,*B*
	Sum of Perimeters	*,*	*,*	*,*	*A,B*	*,*B*	*,*	*A,B*
	Elongation	*,*	*,*	*,*	*,*	*,*	*,*	*,*
	Solidity	*,*	**A**,*	*,*	**A,B**	*,*	*,**B**	*,*
	Closure 100 km	*,*	**A**,*	*,*	*,*	*A*,*	*,*	*,*
	Closure 150 km	*,*	**A,B**	*,*	*,*	*A,B*	*,*	*,*
	Closure 500 km	*,*	**A,B**	*,*	*,*	*A,B*	*,*	*,*
	Dispersion	*,*	*A*,*	*,*	*,*	**A,B**	*,*	*,*B*
	Fragmentation	*,**B**	*A,B*	*,*	*A,B*	*,*	*,*	*,*
	X-Displacement	*,*	*A,B*	*,*	*,*B*	**A**,*	*,*	*,*
	Y-Displacement	*,**B**	*A,**	*,**B**	*,*	**A,B**	*,*	*,*

Table 8: Shape metrics with significant *p*-values (<= 0.05) resulting from the Mann Whitney *U* test performed between the model forecast and Stage IV data for each storm. Significant differences are indicated by an A (HAFS) and B (HWRF-B). Asterisks indicate no difference. The model biases for each storm are indicated by italicized font for smaller values and bold face font for larger values compared with the Stage IV observations.

At the lowest threshold, the HAFS model struggles with underestimating the area and dispersion whereas the HWRF-B model struggles with underestimating the area, dispersion, x-displacement, and sum of perimeters. An underestimation of the x-displacement metric is associated with precipitation objects that are displaced to the west compared with the observations, which may be related to an underlying west track bias in the model forecast

711 (not shown). In general, there are slightly larger track errors for the HWRF-B sample (Figure
712 1a), including a larger west bias compared with HAFS (not shown). These track errors in
713 HWRF-B are consistent with significant x-displacement differences in the location of
714 precipitation. Additionally, for two storms each, there are issues with the HAFS model
715 overestimating the 150-km closure and the HWRF-B model underestimating the
716 fragmentation at the 2 mm hr^{-1} threshold. Both models forecast objects that are too small and
717 too central for their respective samples.

718 At the 5 mm hr^{-1} threshold, the HAFS model consistently forecasts objects that are too
719 empty. Meanwhile, the HWRF-B model consistently forecasts objects that are too central and
720 too displaced to the west, similar to its performance at the 2 mm hr^{-1} threshold. Other issues
721 at this threshold include the HAFS model forecasting objects that are too displaced to the
722 south and the HWRF-B model underestimating the sum of perimeters metric for two storms
723 each. Compared with other thresholds, there is less agreement in the HAFS and HWRF
724 model biases, i.e., there are no similarities in the metrics that are indicated as significant for at
725 least two storms, but again, these models are evaluated for slightly different samples.

726 At the 10 mm hr^{-1} threshold, there are fewer consistent biases in the HAFS model
727 forecast, but there are still problems with the model under-forecasting the sum of perimeters
728 and fragmentation, and over-forecasting the solidity for more than one storm. There is strong
729 agreement in the HWRF-B model underestimating the areas and perimeters and forecasting
730 the objects too far north. Furthermore, the HWRF-B model produces objects that are too full
731 and too displaced to the west for two storms each. The models exhibit the most similar
732 performance at this threshold as they both consistently underestimate the perimeters and
733 fragmentation and overestimated the solidity. Low fragmentation and high solidity are
734 strongly (inversely) related to one another since fragmentation includes solidity in its
735 calculation (Figure 3). Collectively, these two metrics indicate that the convective
736 precipitation is too solid and connected in the model forecast compared with the observation.
737 Since models struggle with capturing the observed precipitation structure in similar ways,
738 there may be a common systematic bias in their representation of convective precipitation in
739 their respective samples.

740 Importantly, the HAFS and HWRF-B models have similar biases in individual storms.
741 For example, the 2 mm hr^{-1} precipitation area is significantly underestimated by both models
742 in four storms: Delta, Isaias, Laura, and Sally. Additionally, both models predict 2 mm hr^{-1}
743 precipitation regions that are too centrally located in Delta, Isaias, and Laura. Precipitation

that is more centrally oriented with respect to the TC center could indicate a positive intensity bias in the model, which is not consistent with the observations (Figure 1b), but caution should be used when linking stratiform precipitation structure with TC intensity. Additionally, these three storms have negative intensity biases in both model forecasts (Table 2). On average, the negative intensity biases in Delta and Laura are smaller compared with the rest of the landfalling cases (Table 2), but again, the dispersion of stratiform precipitation may not be closely linked to TC intensity. Another example is with the Tropical Storm Beta forecast in which both models underestimate the closure at all thresholds and overestimate the dispersion at the 10 mm hr^{-1} threshold. These spatial differences indicate that the model precipitation is too exposed and spread out compared with the observations, which would generally suggest a weaker TC compared with observations. However, the individual TC statistics for Tropical Storm Beta do not suggest a larger negative intensity bias compared with other TCs in the study (Table 2). Other common results include the models under-forecasting the sum of areas metric at all thresholds for Hurricane Sally, over-forecasting closure for Hurricane Zeta at the 2 mm hr^{-1} threshold, and over-forecasting the sum of perimeters metric for Hurricane Hanna at the 5 mm hr^{-1} threshold. These similarities amongst the forecasts of both models suggest that these storms (or, potentially, the GFS initial conditions for these storms) may have characteristics that cause certain metrics to be more difficult to forecast.

For the performance of both models in the maximum precipitation bearing metric, the maximum precipitation is generally forecasted in the correct location, but the model forecast bearings are more like each other than to the observed bearings (Figure 11a-c). Individually, the only storms where there are notable differences between the forecast and observed maximum precipitation are Tropical Storm Beta (Figure 11d-f) and Hurricane Delta (Figure 11g-i). In the HAFS forecast for Tropical Storm Beta, there is a bias for the maximum precipitation to be located too far eastward, while the HWRF-B forecasts has a slight north bias. For Hurricane Delta, both models tend to forecast the greatest precipitation amounts in the north and north-northeast directions from the TC center, but the observed data indicates the maximum precipitation being in the southeast direction for numerous timesteps as well.

These shape metric results indicate that the HAFS model performance varies depending on the rain rate threshold while the HWRF-B model performs similarly across the rain rate thresholds (Table 8). For the HAFS model, persistent issues in forecasting objects that are too small, central, and enclosed within 150 km from the TC center are unique to the 2 mm hr^{-1}

777 threshold; the forecast objects tending to be too empty and fragmented are unique to the 5
778 mm hr^{-1} threshold; and forecast objects being too simple, cohesive, and full for multiple
779 storms is only indicated at the 10 mm hr^{-1} threshold. The HAFS model also performs poorly
780 in the y-displacement for numerous storms across multiple thresholds, and the forecast
781 objects tend to be located too far south at the 2 and 5 mm hr^{-1} thresholds. Track biases for this
782 HAFS sample also indicate a south bias of approximately 10-20 km (not shown), which is
783 consistent with these precipitation results. Additionally, these results indicate that the HAFS
784 model performance differs in the forecasting of stratiform and convective precipitation and
785 emphasize a slight south bias for the forecast objects. This south bias is difficult to compare
786 with an absolute track error (Figure 1a) and might offer additional insight into both
787 precipitation structure and positioning differences compared with the observations.

788 For HWRF-B, the forecast objects tend to be too small, central, cohesive, simple, and
789 displaced to the west for at least two thresholds, and all these spatial biases are present for at
790 least three storms. These persistent biases suggest that there are systematic issues with the
791 HWRF-B precipitation forecasts that are not unique to stratiform or convective precipitation.

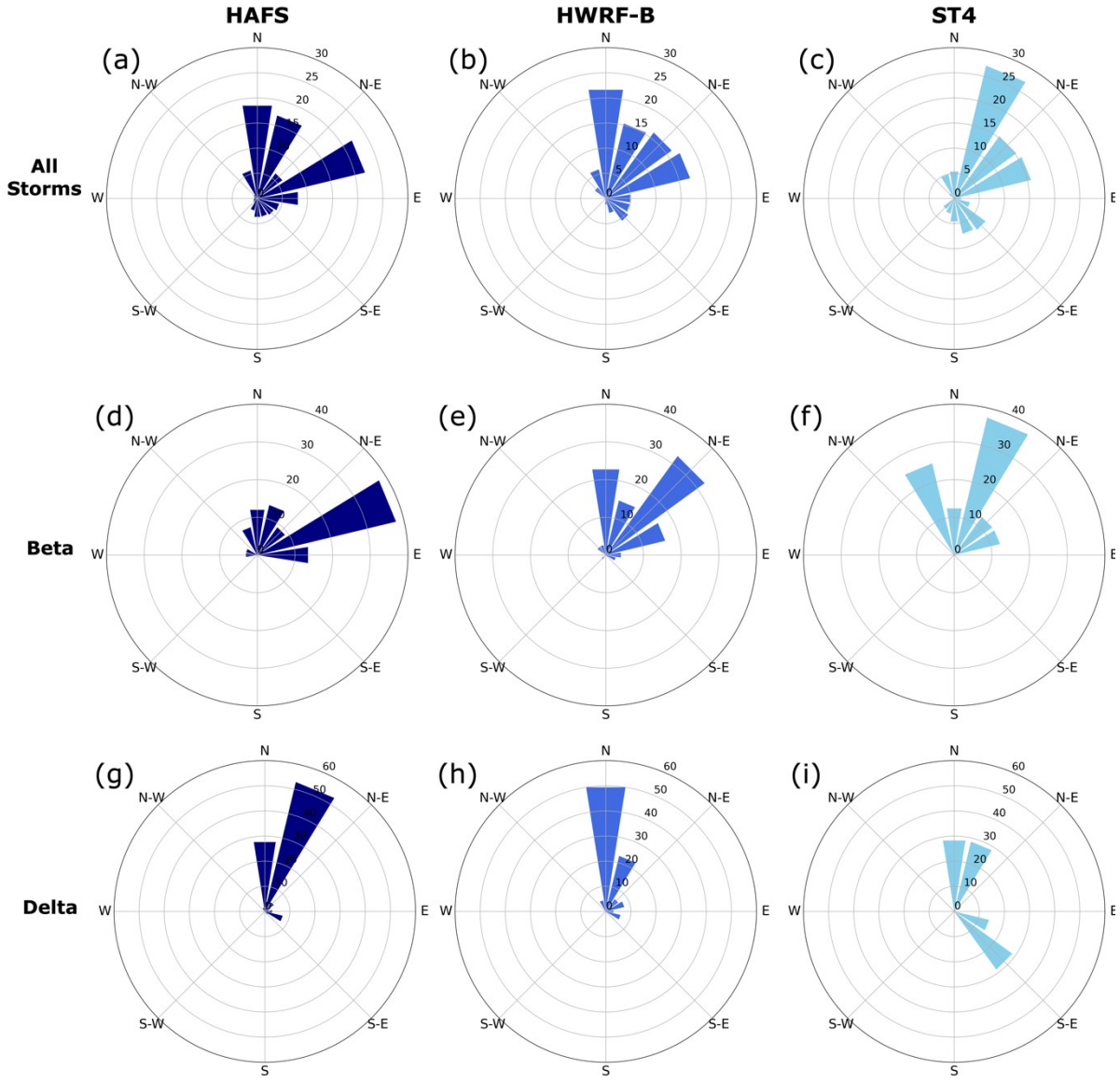

Figure 11: Maximum precipitation bearings with respect to the storm center for (top row) all storms, (middle row) TS Beta, and (bottom row) Hurricane Delta based on precipitation from (left) HAFS, (middle) HWRF-B, and (right) Stage IV (ST4).

Collectively, the precipitation characteristics indicate that both models tend to forecast storms that are too organized and compact compared with the observed storms, especially for the stratiform rain rate regions. For example, the 2 mm hr^{-1} closure for the HAFS model is consistently too high for the inner core region, which indicates a more enclosed stratiform precipitation that is generally consistent with stronger TC circulations and more intense storms (Kieper and Jiang 2012; Matyas and Tang 2019). Precipitation is also too central for the HAFS and HWRF-B models at the lower thresholds, which is a characteristic of stronger storms due to the TCs being more compact. The under-forecasting of fragmentation at the 10

mm hr^{-1} threshold for the HAFS model and at the 2 and 10 mm hr^{-1} thresholds for the HWRF-B model is indicative of more intense storms as well (Zick and Matyas 2016). The exception to the models forecasting more compact storms is the forecast for Tropical Storm Beta in which the model forecast precipitation was more spread and less closed compared with the observations. This also resulted in greater fragmentation in the HAFS forecasts of the storm. Interestingly, the spatial metric results indicate a model bias toward more compact storms, which is generally associated with a positive intensity bias. Instead, there is a negative intensity bias for all storms in the sample for both models (Table 2). These results suggest an inconsistency between the precipitation configuration and the maximum sustained winds in the model forecast.

b. Forecast Verification by Lead Time Evaluation

Generally, the lead time results (Figure 12) indicate spin-up issues associated with the HAFS forecasts and systematic issues with the HWRF-B forecasts. For the HAFS model, at the 2 mm hr^{-1} threshold, the larger number of significant metrics occurring for the first six hours of the model simulations (Figure 12a) may be attributed to the spin-up that occurs when models are initiated via "cold start," as with this version of HAFS. When a model is initialized with fewer observations or with observations that are not representative of the model's resolvable scales, it takes some model integration time for the model to reach a state of balance and spin up some forecast variables such as precipitation. An initial 6-hour spin-up process is suggested based on results for the 2 mm hr^{-1} threshold. Recent and ongoing work has been done to implement initialization and data assimilation in HAFS to mitigate this spin-up issue.

After 54 hours, the model accuracy starts to diminish, suggesting that the model performs well with forecasts up to around two days in advance (excluding the spin-up period). The increasing model error at longer lead times is most evident for the HAFS model forecast of 10 mm hr^{-1} precipitation. The model spin-up has the greatest effect on the HAFS model accuracy at the 2 mm hr^{-1} threshold, but there is a slight suggestion of increasing model errors at longer lead times too. At the 5 mm hr^{-1} threshold, the model is still affected by the spin-up but to a lesser extent, and at the 10 mm hr^{-1} threshold, the spin-up does not seem to have any effect on the model (Figure 12c,e). These results suggest that the HAFS model takes longer to realistically forecast stratiform precipitation (around six hours) compared to convective precipitation. At the 10 mm hr^{-1} threshold, there is around one significant metric for most lead

837 times with increasing numbers of differences at longer lead times (Figure 12e). This is due to
838 the forecast objects being consistently too full at the 10 mm hr^{-1} threshold, which indicates
839 that the model generally did not perform favorably in forecasting the solidity at this
840 threshold. Again, this result points to a HAFS model bias toward more circular, solid
841 convective precipitation structure compared with the observations, and that this model bias is
842 slightly exacerbated at longer lead times.

843 For the HWRF-B model, there is a similar number of significant metrics amongst the lead
844 times at all thresholds (Figure 12b,d,f), which suggests that the HWRF-B forecasts are not
845 affected by the spin-up issue. However, the model consistently under-forecasts the sum of
846 perimeters, dispersion, and fragmentation at the 2 mm hr^{-1} threshold; consistently under-
847 forecasts the sum of perimeters and dispersion at the 5 mm hr^{-1} threshold; and significantly
848 over-forecasts the solidity at the 10 mm hr^{-1} threshold for many lead times. Additionally,
849 there are multiple lead times with significant location metrics at the 2 mm hr^{-1} threshold,
850 which may be attributed to the model forecasting objects that are too displaced to the west
851 (Figure 12b). These results emphasize the systematic issues with the HWRF-B model
852 capturing these metrics, which explains the individual lead times being associated with a
853 similar frequency of significant metrics. For the HWRF-B model, there is no obvious
854 degradation of model forecast skill with lead time (Figure 1b,d,f). This might be attributed to
855 the more advanced data assimilation in this version of the HWRF-B model, or the model
856 degradation with lead time might be obscured by the larger number of significant metrics at
857 all lead times.

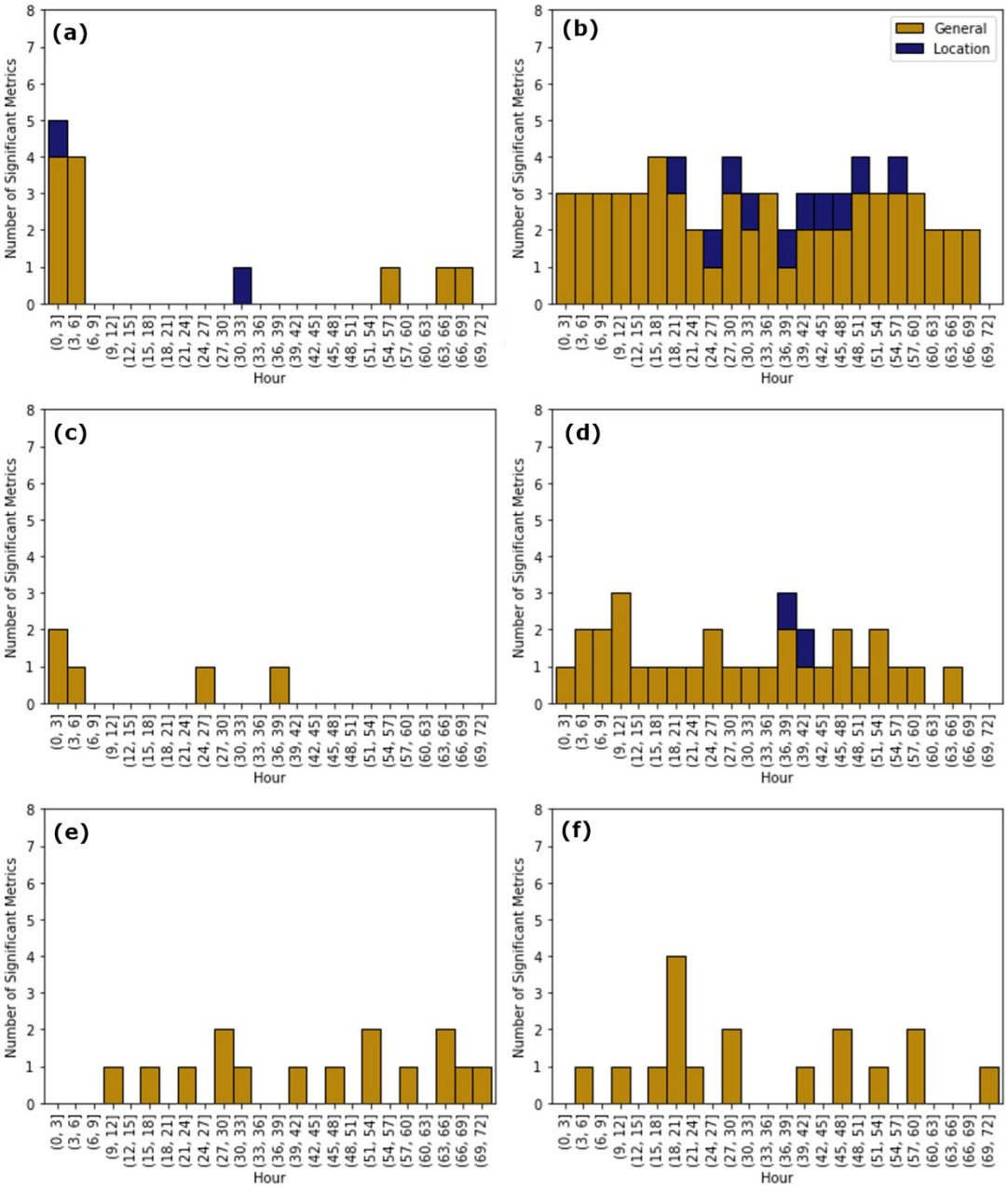

Figure 12: Stacked bar graphs of the number of metrics that are significantly different between the (left) HAFS and Stage IV and (right) the HWRF-B and Stage IV precipitation at each lead time (3-72 hours) at the (top) 2 mm hr^{-1}, (middle) 5 mm hr^{-1}, and (bottom) 10 mm hr^{-1} thresholds. The bars are color-coded based on the category of the metrics.

7. Conclusions

The objectives of this study were to compare point- and object-based precipitation verification methods, evaluate the ability of the HAFS and HWRF-B models to forecast TC precipitation, and assess how model accuracy varies with lead time. In the case study evaluation, the point-based metrics mostly indicated poorer performance in the Hurricane Laura forecast compared with the Hurricane Isaias forecast for both models, whereas the

object-based metrics indicated poorer performance in the Hurricane Isaias forecast. Since the Hurricane Laura forecast from both models had more displacement issues, it was penalized for more misses and false alarms, resulting in lower accuracy. The object-based metrics provided an in-depth spatial analysis of the TC precipitation, and the forecasts were not overly penalized for mismatches in location. These results suggest that point-based methods should be used with caution when evaluating TC precipitation forecasts due to the double penalty problem. Lower precipitation forecast accuracy may be more related to location issues as opposed to precipitation structure issues in the model. Still, since the grid point location is important for forecasting flooding hazards associated with TCs, we recommend that both point- and object-based metrics be assessed in conjunction with one another.

The models were evaluated at three rain rate thresholds: 2 mm hr^{-1} threshold (to evaluate stratiform precipitation) and 5 and 10 mm hr^{-1} thresholds (to evaluate convective precipitation). The results suggest that the HAFS model has separate issues with forecasting stratiform and convective precipitation due to more varied results by rain rate threshold. Specifically, the model forecasted objects that were too small, central, and closed (within 150 km from the TC center) at the 2 mm hr^{-1} threshold; too empty and fragmented at the 5 mm hr^{-1} threshold; and too simple, cohesive, and solid at the 10 mm hr^{-1} threshold. In contrast, the results show that the HWRF-B model has more consistent and perhaps systematic issues at all thresholds including the forecast objects being too small, central, cohesive, simple, and too far west. Collectively, these spatial metrics suggest that both models tend to forecast storms that are too central, closed, and compact compared with the observations, which suggests that the model forecasts may be too intense. This is most evident with closure being too high for the HAFS forecasts, indicating stronger TC circulations, and the dispersion and fragmentation being too low for the forecasts of both models, indicating more organized storms. Instead, there was a negative intensity bias for both models throughout the forecasting period, which suggests that there is an inconsistency between the precipitation configuration and the maximum sustained winds in the model forecast.

The lead time analysis indicated model spin-up issues with the HAFS forecasts and systematic issues with the HWRF-B forecasts. The results showed spin-up issues mostly at the 2 mm hr^{-1} threshold for up to six hours, which suggests that convective precipitation does not seem to be affected by model spin-up whereas stratiform precipitation needs around six hours before it is realistically represented in the HAFS model. Previous studies have suggested that model spin-up issues with precipitation forecasting are predominantly related

902 to convective initiation (Sun et al. 2014), so spin-up issues related to stratiform precipitation
903 are surprising and need to be investigated further. Moreover, the HAFS model performance
904 noticeably worsened after 54 hours at the 2 mm hr^{-1} threshold and there were problems with
905 the model forecasting fuller objects than observed for numerous lead times at the 10 mm hr^{-1}
906 threshold. For the HWRF-B model, object perimeters and dispersion were significantly
907 underestimated for many of the lead times at the 2 and 5 mm hr^{-1} thresholds. Additionally, the
908 fragmentation and the x-displacement were under-forecasted at the 2 mm hr^{-1} threshold, and
909 the solidity was over-forecasted at the 10 mm hr^{-1} threshold across lead times. Like the results
910 of the HWRF-B forecast evaluation by storm, these lead time results imply that there are
911 systematic issues with its forecasts.

912 In this study, the models were expected to have reduced performance during the spin-up
913 period, and this was confirmed with the HAFS model at the 2 mm hr^{-1} threshold. The lead
914 time results did not show spin-up issues with the HWRF-B model. Declining model accuracy
915 at the later lead times, also hypothesized, was shown at 2 and 10 mm hr^{-1} thresholds for the
916 HAFS model, but this was not indicated with the HWRF-B forecasts, which instead showed
917 systematic biases across all lead times through 72 hours. This result for HWRF-B might be
918 due to its more advanced data assimilation and vortex initialization. Another possibility is
919 that the decreasing performance with lead time might be obscured by the larger number of
920 significant structural differences at all lead times.

921 There were several limitations in this study including (1) the sensitivity of the shape
922 metrics to the search radius, (2) the small sample size, and (3) errors related to the Mann
923 Whitney *U* test. First, precipitation from non-tropical systems is occasionally included in the
924 precipitation object if it is located within the search radius, which can cause inaccuracies.
925 Visual inspection of the output found that this was only a minor factor in Hurricane Zeta.
926 Second, this study could provide stronger evidence of systematic issues with the models if a
927 greater sample of storms were analyzed. Third, there are limitations with the Mann Whitney
928 *U* test as it only distinguishes differences in the shape metrics based on the medians if the
929 distributions are equal. When distributions are not equal, a significant result indicates
930 differences in the distributions more generally, which is not as robust as a comparison of the
931 medians. To confirm that the Mann Whitney *U* test was detecting actual differences in the
932 median, we performed post-hoc comparisons for all the results reported in this study. Lastly,
933 since the Mann Whitney *U* test is performed many times in this analysis, the possibility of

type I errors (significant differences are detected when there are none) and type II errors (no significant differences are detected when there are actual differences) is increased.

This study demonstrated the usefulness of implementing an object-based method for model verification to thoroughly assess the HAFS and HWRF-B precipitation forecasts using a variety of spatial metrics. Weaknesses in the HAFS and HWRF-B precipitation forecasts at three rain rate thresholds were identified, which will support model developers to work towards improving these flaws, through improvements to the microphysics, convective parameterization, or other model physics and dynamics. Forecasters seeking guidance from these models can also use this research to compensate for the model deficiencies that were detected with regards to forecasting TC precipitation. This study contributes to existing research on the HAFS and HWRF-B models by evaluating their ability to forecast TC precipitation structures rather than only the track or intensity. This research also evaluates more recent versions of the models (2020) compared to previous studies that analyze the model performance during earlier years. The results suggest that future studies are needed to investigate spin-up issues of stratiform precipitation in the HAFS model, the systematic issues in the HWRF-B model, and the positive bias in forecasting the storm intensity for both models. Lastly, future studies also need to investigate the inconsistency between model forecast precipitation structure and TC intensity.

Acknowledgments.

This work is funded by the National Science Foundation (AGS- 2011981). We also received funding from the Virginia Tech Graduate School through a Dean's Diversity Assistantship. We appreciate the thoughtful comments and suggestion from three anonymous reviewers.

Disclaimer: 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.

Software and data supporting this study are available at https://dx.doi.org/10.7294/21737039. Additionally, we use datasets that were derived from

965 the following public domain resources: (1) All Stage IV precipitation data are openly
966 available from the NCAR Earth Observing Laboratory archive at
967 https://doi.org/10.5065/D6PG1QDD as cited in Lin and Mitchell (2005) and (2) HURDAT2
968 data are openly available from the National Hurricane Center https://www.nhc.noaa.gov/data/
969 as cited in Landsea and Franklin (2013).

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Alaka, G. J., 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. Meteorol. Soc.*, **103**, E680–E703, https://doi.org/10.1175/BAMS-D-20-0134.1.

Biswas, M., L. Carson, K. Newman, D. Stark, E. Kalina, E. Grell, and J. Frimel, 2018: *Community HWRF Users' Guide V4.0a*. https://dtcenter.org/sites/default/files/community-code/hwrf/docs/users_guide/HWRF-UG-2018.pdf.

Chu, Q., Z. Xu, Y. Chen, and D. Han, 2018: Evaluation of the ability of the Weather Research and Forecasting model to reproduce a sub-daily extreme rainfall event in Beijing, China using different domain configurations and spin-up times. *Hydrology and Earth System Sciences*, **22**, 3391–3407, https://doi.org/10.5194/hess-22-3391-2018.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

Dong, J., and Coauthors, 2020: The Evaluation of Real-Time Hurricane Analysis and Forecast System (HAFS) Stand-Alone Regional (SAR) Model Performance for the 2019 Atlantic Hurricane Season. *Atmosphere*, **11**, 617, https://doi.org/10.3390/atmos11060617.

Dvorak, V. F., 1990: *A workbook on tropical clouds and cloud systems observed in satellite imagery*. U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service [and] National Weather Service, 244 pp.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, **239**, 179–202, https://doi.org/10.1016/S0022-1694(00)00343-7.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying Forecasts Spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1373, https://doi.org/10.1175/2010BAMS2819.1.

Gopalakrishnan, S. and Coauthors, 2021: *2020 HFIP R&D Activities Summary: Recent Results and Operational Implementation*. National Oceanic and Atmospheric Administration,.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q.J.R. Meteorol. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Harris, L. M., and S.-J. Lin, 2013: A Two-Way Nested Global-Regional Dynamical Core on the Cubed-Sphere Grid. *Monthly Weather Review*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

Hazelton, A., and Coauthors, 2021a: 2019 Atlantic Hurricane Forecasts from the Global-Nested Hurricane Analysis and Forecast System: Composite Statistics and Key Events. *Weather and Forecasting*, **36**, 519–538, https://doi.org/10.1175/WAF-D-20-0044.1.

——, G. J. Alaka, L. Cowan, M. Fischer, and S. Gopalakrishnan, 2021b: 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, 2022a: Performance of 2020 Real-Time Atlantic Hurricane Forecasts from High-Resolution Global-Nested Hurricane Models: HAFS-globalnest and GFDL T-SHiELD. *Weather and Forecasting*, **37**, 143–161, https://doi.org/10.1175/WAF-D-21-0102.1.

——, S. Gopalakrishnan, and J. A. Zhang, 2022b: Comparison of The Hybrid EDMF and Modified EDMF-TKE PBL Schemes in 2020 Tropical Cyclone Forecasts from the Global-nested Hurricane Analysis and Forecast System. *Weather and Forecasting*, **37**, 457–476.

Hazelton, A. T., L. Harris, and S.-J. Lin, 2018: Evaluation of Tropical Cyclone Structure Forecasts in a High-Resolution Version of the Multiscale GFDL fvGFS Model. *Wea. Forecasting*, **33**, 419–442, https://doi.org/10.1175/WAF-D-17-0140.1.

Hernández Ayala, J. J., and C. J. Matyas, 2016: Tropical cyclone rainfall over Puerto Rico and its relations to environmental and storm-specific factors. *Int. J. Climatol.*, **36**, 2223–2237, https://doi.org/10.1002/joc.4490.

Hohenegger, C., and C. Schar, 2007: Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales. *Bulletin of the American Meteorological Society*, **88**, 1783–1794, https://doi.org/10.1175/BAMS-88-11-1783.

Kieper, M. E., and H. Jiang, 2012: Predicting tropical cyclone rapid intensification using the 37 GHz ring pattern identified from passive microwave measurements. *Geophys. Res. Lett.*, **39**, L13804, https://doi.org/10.1029/2012GL052115.

Kirkland, J. L., and S. E. Zick, 2019: Regional Differences in the Spatial Patterns of North Atlantic Tropical Cyclone Rainbands Through Landfall. *Southeastern Geographer*, **59**, 294–320, https://doi.org/10.1353/sgo.2019.0023.

1049 Landsea, C. W., and J. L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and
1050 Presentation of a New Database Format. *Mon. Wea. Rev.*, **141**, 3576–3592,
1051 https://doi.org/10.1175/MWR-D-12-00254.1.

1052 Latto, A., A. Hagen, and R. Berg, 2021: *Tropical cyclone report: Hurricane Isaias. 30 July -*
1053 *4 August 2020.* National Hurricane Center,.

1054 Lau, K. M., and H. T. Wu, 2003: Warm rain processes over tropical oceans and climate
1055 implications. *Geophysical Research Letters*, **30**,
1056 https://doi.org/10.1029/2003GL018567.

1057 Li, J., K.-L. Hsu, A. AghaKouchak, and S. Sorooshian, 2016: Object-Based Assessment of
1058 Satellite Precipitation Products. *Remote Sensing*, **8**, 547,
1059 https://doi.org/10.3390/rs8070547.

1060 Lin, Y., and K. E. Mitchell, 2005: 1.2 the NCEP stage II/IV hourly precipitation analyses:
1061 Development and applications. 19th Conf. Hydrology, Citeseer.

1062 Liu, Q., and Coauthors, 2020: Vortex Initialization in the NCEP Operational Hurricane
1063 Models. *Atmosphere*, **11**, 968, https://doi.org/10.3390/atmos11090968.

1064 Lonfat, M., F. Marks, and S. S. Chen, 2004: Precipitation distribution in tropical cyclones
1065 using the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager: A global
1066 perspective. *Monthly Weather Review*, **132**, 1645–1660, https://doi.org/10.1175/1520-
1067 0493(2004)1322.0.CO;2.

1068 Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, **20**, 130–141,
1069 https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

1070 Luitel, B., G. Villarini, and G. A. Vecchi, 2016: Verification of the skill of numerical weather
1071 prediction models in forecasting rainfall from U.S. landfalling tropical cyclones.
1072 *Journal of Hydrology*, https://doi.org/10.1016/j.jhydrol.2016.09.019.

1073 Matyas, C. J., 2007: Quantifying the shapes of U.S. landfalling tropical cyclone rain shields.
1074 *The Professional Geographer*, **59**, 158–172, https://doi.org/10.1111/j.1467-
1075 9272.2007.00604.x.

1076 ——, and J. Tang, 2019: Measuring Radial and Tangential Changes in Tropical Cyclone Rain
1077 Fields Using Metrics of Dispersion and Closure. *Advances in Meteorology*, **2019**, 14,
1078 https://doi.org/10.1155/2019/8613943.

1079 ——, S. E. Zick, and J. Tang, 2018: Using an Object-Based Approach to Quantify the Spatial
1080 Structure of Reflectivity Regions in Hurricane Isabel (2003). Part I: Comparisons
1081 between Radar Observations and Model Simulations. *Mon. Wea. Rev.*, **146**, 1319–
1082 1340, https://doi.org/10.1175/MWR-D-17-0077.1.

1083 McBride, J. L., and E. E. Ebert, 2000: Verification of Quantitative Precipitation Forecasts
1084 from Operational Numerical Weather Prediction Models over Australia. *Weather and*
1085 *Forecasting*, **15**, 103–121, https://doi.org/10.1175/1520-
1086 0434(2000)015<0103:VOQPFF>2.0.CO;2.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2015: Assessment and Implications of NCEP Stage IV Quantitative Precipitation Estimates for Product Intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

NOAA, 2020: Next Generation Weather Radar (NEXRAD). *National Centers for Environmental Information (NCEI)*,. https://www.ncei.noaa.gov/products/radar/next-generation-weather-radar (Accessed December 2, 2022).

Pasch, R. J., R. Berg, D. Roberts, and P. Papin, 2021: *Tropical cyclone report: Hurricane Laura. 20-29 August 2020.* National Hurricane Center,.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh Model's Ability to Predict Mesoscale Convective Systems Using Object-Based Evaluation. *Wea. Forecasting*, **30**, 892–913, https://doi.org/10.1175/WAF-D-14-00118.1.

Rappaport, E., 2014: Fatalities in the United States from Atlantic tropical cyclones: New data and interpretation. *Bull. Amer. Meteor. Soc.*, **95**, 341–346, https://doi.org/10.1175/BAMS-D-12-00074.1.

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. Michaelides, Ed., Springer, 419–452.

Roth, D., 2021: Tropical Cyclone Rainfall Data. *Tropical Cyclone Rainfall*,. https://www.wpc.ncep.noaa.gov/tropical/rain/tcrainfall.html (Accessed October 20, 2022).

Schaefer, J. T., 1990: The Critical Success Index as an Indicator of Warning Skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Schumacher, C., and R. A. Houze, 2003: Stratiform rain in the tropics as seen by the TRMM precipitation radar. *Journal of Climate*, **16**, https://doi.org/10.1175/1520-0442(2003)016<1739:sritta>2.0.co;2.

Sierra-Lorenzo, M., J. Medina, J. Sille, A. Fuentes-Barrios, S. Alfonso-Águila, and T. Gascon, 2022: Verification by Multiple Methods of Precipitation Forecast from HDRFFGS and SisPI Tools during the Impact of the Tropical Storm Isaias over the Dominican Republic. *Atmosphere*, **13**, 495, https://doi.org/10.3390/atmos13030495.

Spearman, C., 1961: *"General Intelligence" Objectively Determined and Measured*. Appleton-Century-Crofts, 59 pp.

Sun, J., and Coauthors, 2014: USE OF NWP FOR NOWCASTING CONVECTIVE PRECIPITATION: Recent Progress and Challenges. *Bulletin of the American Meteorological Society*, **95**, 409–426.

Tokay, A., D. A. Short, C. R. Williams, W. L. Ecklund, and K. S. Gage, 1999: Tropical Rainfall Associated with Convective and Stratiform Clouds: Intercomparison of Disdrometer and Profiler Measurements. *Journal of Applied Meteorology and*

Climatology, **38**, 302–320, https://doi.org/10.1175/1520-0450(1999)038<0302:TRAWCA>2.0.CO;2.

Tuleya, R., M. DeMaria, and R. Kuligowski, 2007: Evaluation of GFDL and simple statistical model rainfall forecasts for US landfalling tropical storms. *Weather and Forecasting*, **22**, 56–70, https://doi.org/10.1175/WAF972.1.

Unified Forecast System, What is UFS? *Unified Forecast System*,. https://ufscommunity.org/about/what-is-ufs/ (Accessed October 20, 2022).

Velden, C. S., and Coauthors, 2006: The Dvorak Tropical Cyclone Intensity Estimation Technique: A Satellite-Based Method that Has Endured for over 30 Years. *Bull. Amer. Meteor. Soc.*, **87**, 1195–1210, https://doi.org/10.1175/BAMS-87-9-1195.

Villarini, G., J. A. Smith, M. L. Baeck, T. Marchok, and G. A. Vecchi, 2011: Characterization of rainfall distribution and flooding associated with US landfalling tropical cyclones: Analyses of Hurricanes Frances, Ivan, and Jeanne (2004). *Journal of Geophysical Research-Atmospheres*, **116**, https://doi.org/10.1029/2011jd016175.

Villarini, G., W. Zhang, P. Miller, D. R. Johnson, L. E. Grimley, and H. J. Roberts, 2022: Probabilistic rainfall generator for tropical cyclones affecting Louisiana. *International Journal of Climatology*, **42**, 1789–1802, https://doi.org/10.1002/joc.7335.

Wilcoxon, F., 1992: Individual Comparisons by Ranking Methods. *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N.L. Johnson, Eds., *Springer Series in Statistics*, Springer, 196–202.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences, Volume 100, Third Edition*. 3 edition. Academic Press, 704 pp.

Wong, M., and W. C. Skamarock, 2016: Spectral Characteristics of Convective-Scale Precipitation Observations and Forecasts. *Monthly Weather Review*, **144**, 4183–4196, https://doi.org/10.1175/MWR-D-16-0183.1.

Zhang, X., S. G. Gopalakrishnan, S. Trahan, T. S. Quirino, Q. Liu, Z. Zhang, G. Alaka, and V. Tallapragada, 2016: Representing Multiple Scales in the Hurricane Weather Research and Forecasting Modeling System: Design of Multiple Sets of Movable Multilevel Nesting and the Basin-Scale HWRF Forecast Application. *Wea. Forecasting*, **31**, 2019–2034, https://doi.org/10.1175/WAF-D-16-0087.1.

Zick, S., C. Matyas, G. Lackmann, J. Tang, and B. Bennett, 2022: Illustration of an object-based approach to identify structural differences in tropical cyclone wind fields. *Quarterly Journal of the Royal Meteorological Society*, **148**, 2587–2606, https://doi.org/10.1002/qj.4326.

Zick, S. E., 2020: Quantifying Extreme Precipitation Forecasting Skill in High-Resolution Models Using Spatial Patterns: A Case Study of the 2016 and 2018 Ellicott City Floods. *Atmosphere*, **11**, 136, https://doi.org/10.3390/atmos11020136.

——, and C. J. Matyas, 2016: A Shape Metric Methodology for Studying the Evolving Geometries of Synoptic-Scale Precipitation Patterns in Tropical Cyclones. *Annals of*

the American Association of Geographers, **106**, 1217–1235, https://doi.org/10.1080/24694452.2016.1206460.