Evaluation of the impact of different microphysics scheme on HAFS model microphysics forecasts using GOES-R infrared images

A DTC-visitor project

Shaowu Bao, Coastal Carolina University

Summary

This study evaluates the performance of two configurations of the Hurricane Analysis and Forecast System (HAFS), named HFSA and HFSB. While these configurations have several differences, one key distinction lies in their microphysics parameterization schemes. Using infrared brightness temperature data from GOES-R satellite observations as a proxy for cloud and hydrometeor characteristics, the study assesses the models' skill in simulating three 2023 Atlantic hurricanes: Idalia, Lee, and Ophelia. Statistical metrics including probability density functions, composite images, target diagrams, Taylor diagrams, Fractions Skill Score, and Equitable Threat Score are employed to compare the models' performance across various thresholds and forecast lengths. The results consistently demonstrate that HFSA outperforms HFSB, exhibiting lower biases, higher correlation coefficients, and better predictive skill. Both models capture the vortex structures and asymmetries of the hurricanes well. However, they tend to overestimate cloud coverage and height compared to observations, with HFSB showing a more pronounced bias. Despite HFSA's superior performance, there remains room for improvement in both models' forecasting skill, particularly at longer lead times and lower thresholds. The study highlights the importance of evaluating and refining microphysics parameterizations to enhance tropical cyclone prediction capabilities. Future research should focus on expanding the analysis to a larger sample of storms, exploring additional metrics, and conducting targeted diagnostics to guide the development of improved hurricane forecasting tools.

Background and motivation

The National Oceanic and Atmospheric Administration (NOAA) has developed the Hurricane Analysis and Forecast System (HAFS), a next-generation system designed to improve tropical cyclone prediction capabilities. At the time this project was initiated, two configurations of HAFS were being considered for operational implementation in 2023, differing primarily in their microphysics parameterization schemes. Since then, HAFS has been adopted into NOAA’s operational modeling suite and has been in use for over a year, replacing previous hurricane forecast models.

While HAFS has demonstrated improved track and intensity forecasting over existing models (Dong et al. 2020; Zhang et al. 2022; Bao et al. 2022), its ability to accurately predict precipitation from tropical cyclones remains unclear. Flooding from extreme rainfall is a major hazard, causing over a quarter of tropical cyclone deaths. However, model validation efforts have primarily focused on track and intensity rather than precipitation forecasts. This is problematic because rainfall impacts can occur far from a storm's center, even with accurate track predictions (Lonfat et al. 2007; Marchok et al. 2007).

Regional hurricane models have shown better quantitative precipitation forecasting (QPF) skill than global models (Ko et al. 2020), underscoring the importance of evaluating HAFS's QPF performance. Previous analysis of an early HAFS version found an underestimation of lower rainfall percentiles and

overestimation of higher percentiles (Green et al. 2022). To address this knowledge gap, the Developmental Testbed Center is undertaking a QPF evaluation of the two proposed HAFS configurations for operational implementation.

The contrasting microphysics parameterizations are expected to significantly influence QPF skill differences between the configurations. A key distinction is that HFSA only has prognostic mixing ratios, while HFSB also includes prognostic variables for cloud properties, precipitation characteristics, and aerosols. Condensation, evaporation, and precipitation processes also differ between the schemes. Understanding how these microphysics differences impact QPF forecast skill is critical for selecting the optimal HAFS configuration.

To diagnose this, we first need to evaluate the models' skills related to microphysics by comparing two different microphysics schemes with observations to see how well they model hydrometeors. However, since direct observations of hydrometeors are rare, we will convert the model forecasts into synthetic GOES satellite infrared channel images and compare them with actual GOES images.

Goal and Objectives

The study uses remote sensing and atmospheric reanalysis data, as well as diagnostic numerical experiments, to assess the effects of the GFDL and Thompson microphysics schemes on the QPF skills of the two HAFS configurations and make recommendations to model developers.

Data

1. Storm cases

The study examined three 2023 Atlantic hurricanes: Hurricane Lee (September 5, 12:00 PM - September 18, 3:00 PM), Hurricane Idalia (August 26, 6:00 PM - September 8, 12:00 AM), and Hurricane Ophelia (September 21, 12:00 PM - September 24, 6:00 PM). The tracks of these hurricanes are shown in Figure 1. Note all times in this report are expressed in Coordinated Universal Time (UTC)

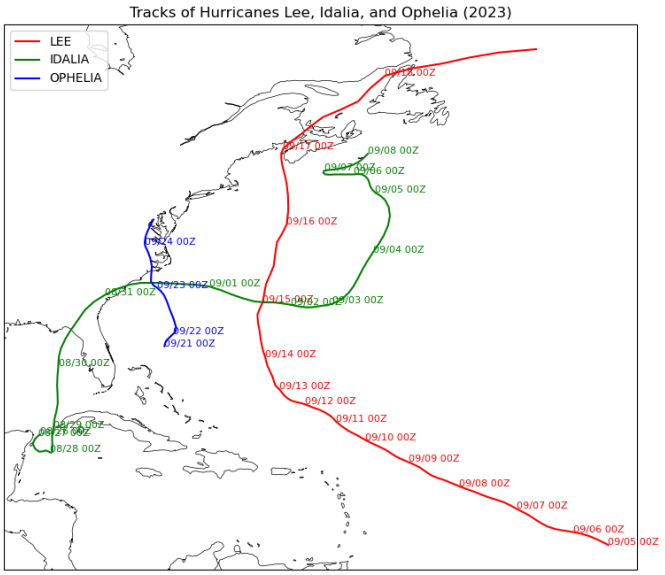

Figure 1 Tracks of the three hurricanes evaluated in the study

2. GOES observation

The study utilized GOES-R infrared observations at a wavelength of 10.3 µm (the "Clean" longwave window) Advanced Baseline Imager (ABI) images in full-disk mode. GOES-R, operated by NOAA, is a series of geostationary weather satellites designed for advanced imaging and atmospheric measurements. It captures images every 5 minutes with a resolution ranging from 3 to 5 km (Source: GOES-R ABI Bands Table (GOES-R) and GOES Rebroadcast (NOAAasis)).

3. CRTM and synthetic images

The Community Radiative Transfer Model (CRTM) is a tool used in atmospheric sciences to simulate the interaction of radiation with the Earth's atmosphere. It converts model simulation data into synthetic satellite images, including those from GOES-R. By generating synthetic images for various channels, such as infrared, CRTM facilitates direct comparisons with observed GOES-R images, aiding in the validation and refinement of simulation models.

The simulation data required for CRTM includes atmospheric temperature, moisture profiles, surface properties, and hydrometeor characteristics like clouds, ice, snow, and rainwater. These parameters are crucial for accurately replicating the radiative properties observed by satellites.

By comparing the synthetic GOES images with observed ones, we can evaluate the model's ability to simulate hydrometeors. This comparison provides insights into the model's accuracy in representing atmospheric elements, enabling further refinement of simulation processes and improving predictive capabilities.

The GOES-R data were downloaded from AWS S3 buckets using the s3fs tool. The GOES-R data were spatially interpolated onto selected points of the model's high-resolution, storm-following output grid using a spline algorithm. The analysis was restricted to the region within a 6-degree radius of the storm center. Accordingly, both the input GOES-R observations and the target model grid points (defined by

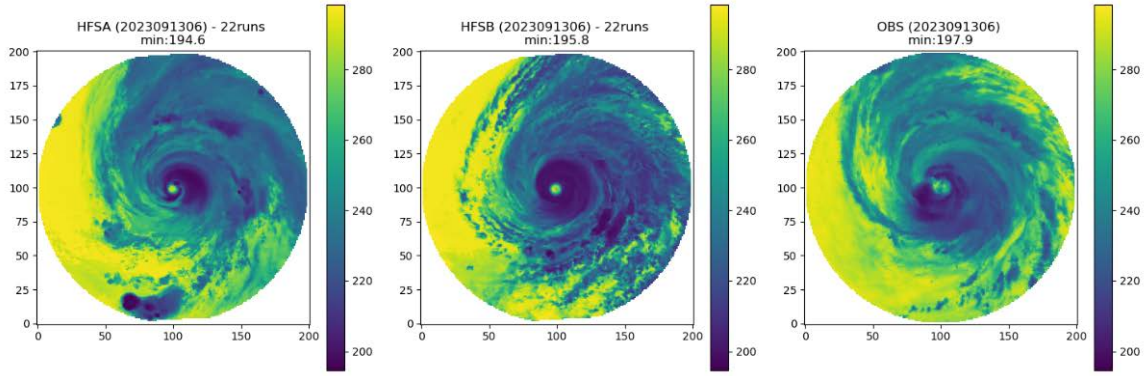

Figure 2 Hurricane Lee at the valid time of 2023-09-13 06Z. The three panels show the brightness temperature for HFSA (left), HFSB (middle) and the Observed (right). The color bar scale ranges from 200 K to 280 K, where warmer colors (yellow) indicate cloud-free areas and cooler colors (purple) represent regions with clouds and hydrometeors. The tick mark labels are grid indices.

latitude/longitude coordinates) were limited to this 6-degree radius; data outside this region were excluded prior to interpolation. The storm center corresponds to the center of the model's storm-following domain. The intervals for latitude (Δlat) and longitude (Δlon) are set at 0.06 degrees, creating a 200x200 grid. An example of this re-gridded data is shown in Figure 2

Statistics methods and Results

The most direct way to compare model and observed satellite images is to visually examine each individual image. However, this method is time-consuming, and the model simulations contain random uncertainties (noise) that make it difficult to draw conclusions about the model's systematic performance and skill evaluation.

Instead, we rely on statistical metrics for the evaluation, including probability density functions, target diagrams, Taylor diagrams, Fractions Skill Score, and Equitable Threat Score. Additionally, we use composite images, which help to mitigate random errors, revealing systematic biases. Below, we present the results of these statistical and composite evaluations.

1. Probability Density Function (PDF)

- **Hurricanes Idalia and Ophelia**

The observed brightness temperatures exhibit a sharp peak around 285 K, suggesting that the majority of the observed brightness temperatures fall within this range, indicative of clear, non-cloudy areas. The peak at around 285K suggests that the storm size is relatively small and a considerable portion of the domain remains cloud free during the evolution of the storm .

Both model simulations show peaks around the same brightness temperature region but differ in distribution shape and magnitude. HFSB exhibits notable deviations from observations at both the lower and higher temperature ranges. These differences are consistent with an overrepresentation of colder brightness temperatures, likely due to excessive or overly extensive cold cloud tops in the simulation. This leads to a corresponding underrepresentation in the warmer range, as the total probability must remain conserved. In contrast, HFSA aligns more closely with the observed distribution, particularly in the warmer temperature range, indicating a more balanced representation of cloud and surface temperatures.

The models both show a broader distribution and higher frequency in the lower temperature range (around 200-220 K), suggesting that both HFSA and HFSB forecast more cloud and hydrometeor coverage than observed in the actual GOES data. The observed PDF for Ophelia exhibits a higher frequency of mid-range BTs (~220-260 K) compared to Idalia, suggesting differences in cloud structure, such as more extensive anvil or mid-level clouds

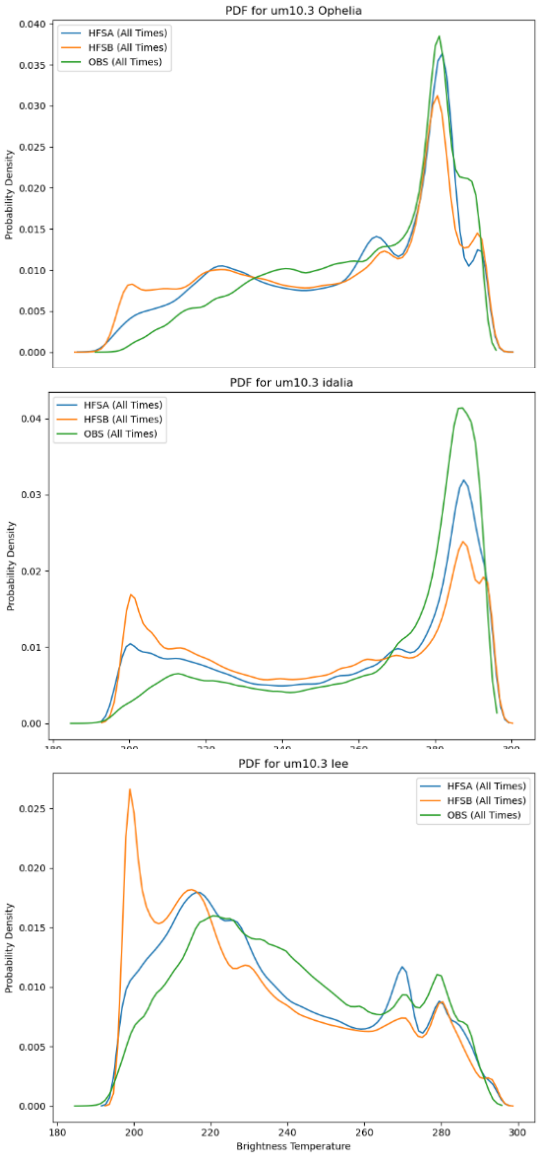

Figure 3 PDF for hurricanes Ophelia (upper), Idalia (middle) and Lee (bottom)

- **Hurricane Lee**

The observed brightness temperatures exhibit a prominent peak around 220 K, suggesting a relatively large vortex with clouds and hydrometers and a small fraction of clear, non-cloudy areas.

The models both show higher frequency in the lower temperature range (around 200-230 K) compared to the observed data, suggesting that both HFSA and HFSB overestimated colder cloud and hydrometeor coverage than observed in the actual GOES data, and thus underestimated the cloud-free areas, a pattern consistent with the one shown in Idalia and Ophelia.

Overall, the PDF figures indicate that both HFSA and HFSB overestimate cloud coverage and the height of cloud tops, resulting in larger regions of cold temperatures. This discrepancy is more pronounced in HFSB than in HFSA.

2. Composite images

Figure 4 shows composite infrared brightness temperature plots for HFSA, HFSB, and OBS for hurricanes Idalia, Lee, and Ophelia. The overall color patterns and distributions are quite similar among HFSA, HFSB, and OBS, indicating that both models are reasonably effective at matching the general pattern of the observed brightness temperature data. While the observed large-scale thermal structures differ notably between the three storms shown in Figure 4, both the HFSA and HFSB models generally succeed in capturing the spatial distribution of the key warm (bright yellow/orange) and cold (dark blue) features

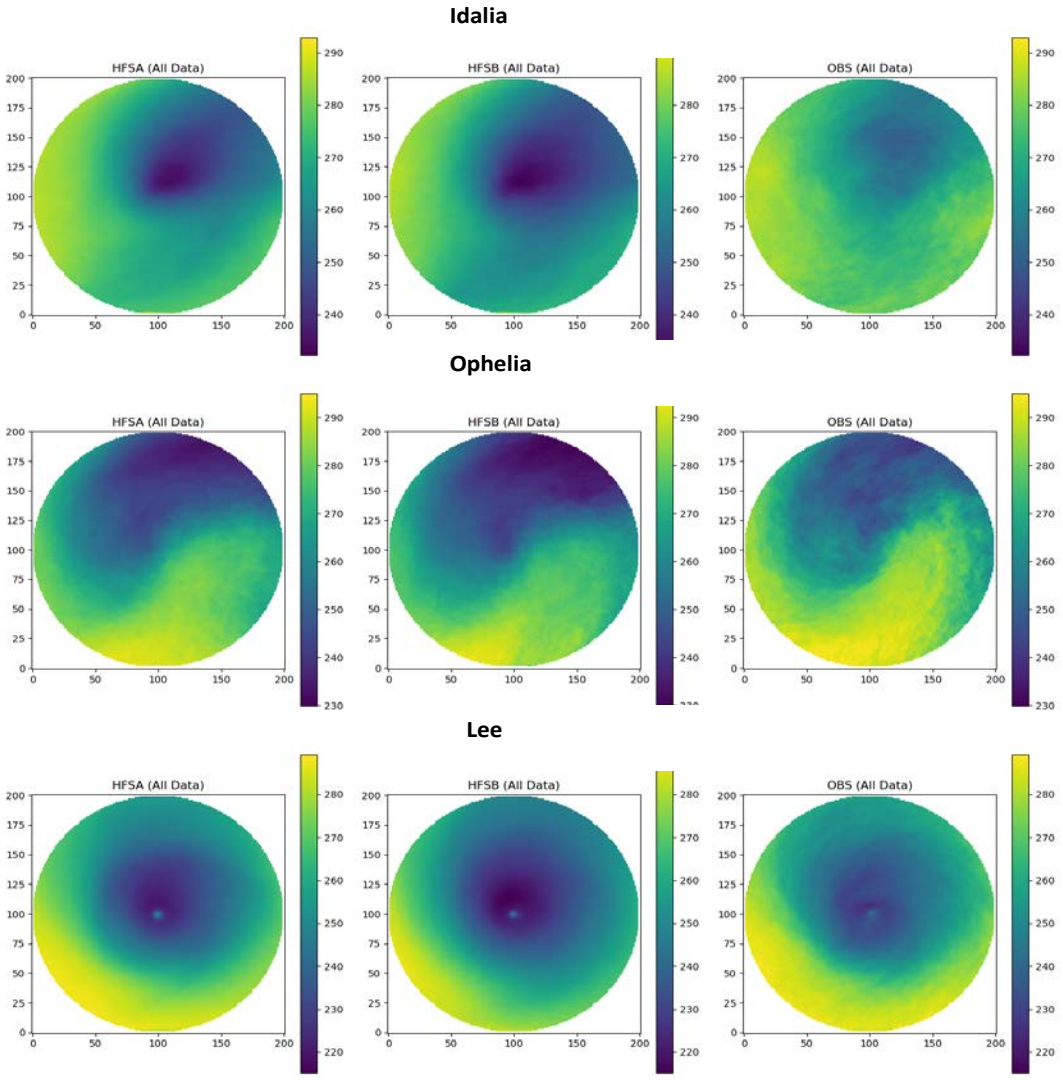

Figure 4 Composite images using HFSA, HFSB and OBS for hurricanes Idalia (1242 forecasts and 98 observed), Ophelia (180 forecasts and 27 observed), and Lee (1763 forecasts and 105 observed)

seen in the corresponding observations for each case. For example, the cold temperature areas are located in the northeastern quadrant for Idalia, the northwestern quadrant for Ophelia, and the northern half for Lee in all HFSA, HFSB, and OBS plots, suggesting that both models captured the asymmetric structures of the storms well.

However, upon closer inspection, the HFSB plot (center panels) appears noticeably darker compared to the OBS plot, indicating a bias towards overestimating colder brightness temperatures. The HFSA plot (left hand panels) also looks slightly darker than OBS, but the difference is subtler. This suggests that both models, particularly HFSB, may systematically overestimate the coldness of the brightness temperatures, indicating higher, colder clouds/hydrometeors compared to actual observations.

This conclusion aligns with the PDF evaluations, which show that both HFSA and HFSB overestimate cloud coverage and higher cloud tops, leading to larger regions of cold temperatures. This discrepancy is

more evident for HFSB than HFSA. Additionally, the HFSA and HFSB plots appear smoother, with more gradual color gradients compared to the patchier, more granular appearance of OBS. This smoothing effect results from averaging multiple model forecasts with different initialization times and forecast lengths. In contrast, for each valid time, there is only one observed GOES-R image. For Idalia, HFSA and HFSB are averaged from 1242 forecasts with 98 unique valid times, for Lee from 1763 forecasts with 105 unique valid times, and for Ophelia from 180 forecasts with 27 unique valid times.

3. Composites by valid time

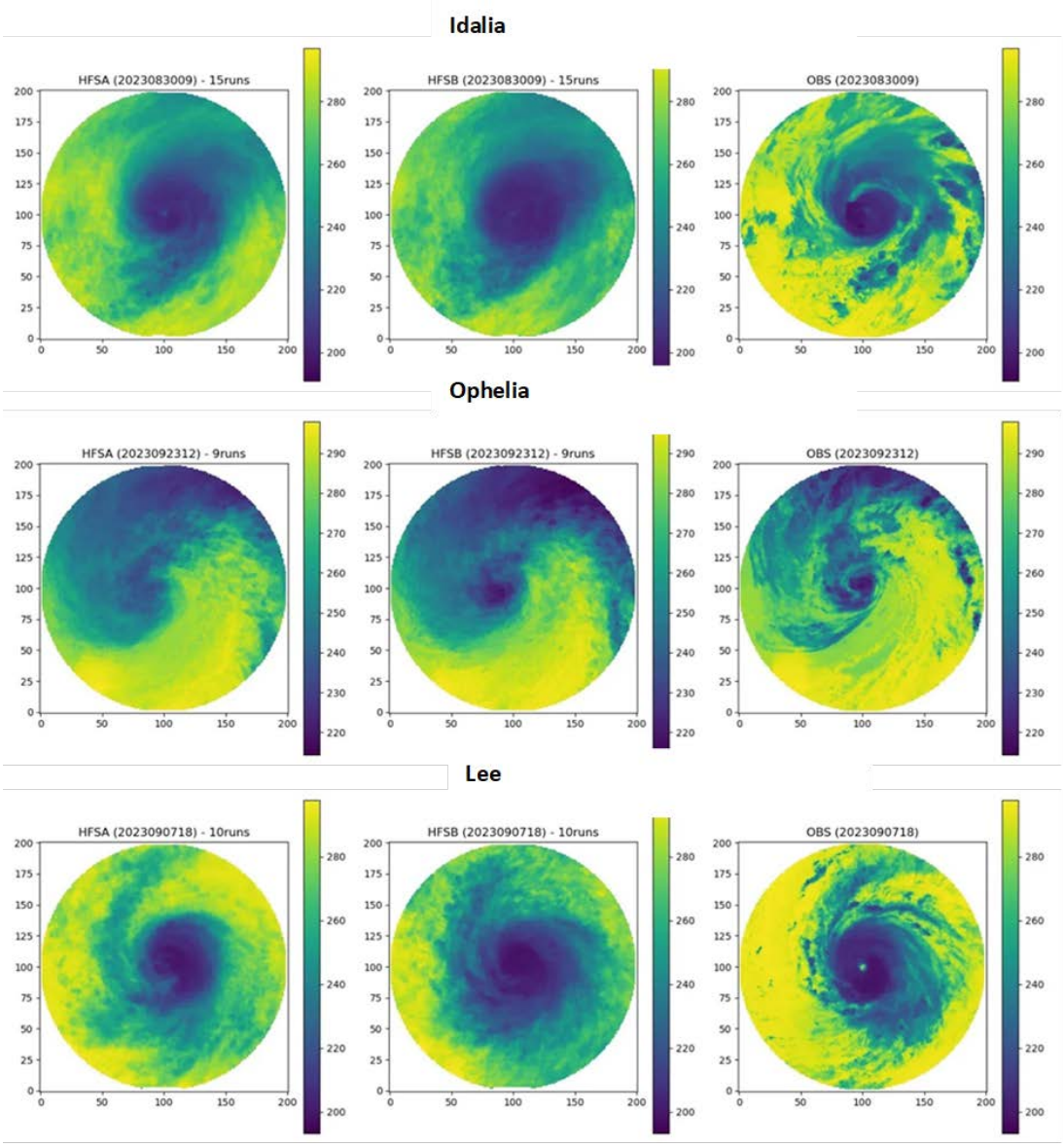

Figure 5 Infrared brightness temperature plots for Hurricanes Idalia, Ophelia, and Lee comparing HFSA, HFSB (multiple forecasts composites), and observed GOES-R (single observation) at specific valid times. The valid times are chosen when the hurricane vortex is relatively mature, well-structured, and at high intensity.

Comparing the individual valid-time plots (Fig. 5), the overall color patterns and distributions remain fairly consistent with the composite image analysis. The HFSA, HFSB, and OBS plots show broadly similar

locations of the warmest (bright yellow/green) and coldest (dark blue/purple) temperature regions for each storm. This indicates that both models capture the overall storm structures reasonably well at these specific times.

However, some differences are more apparent in these individual plots. The HFSB images appear considerably darker and colder compared to the corresponding OBS images for all three storms. The HFSA images are also slightly darker than OBS but to a lesser extent than HFSB. This supports the previous conclusion that both models, particularly HFSB, tend to overestimate the extent and intensity of cold brightness temperatures, likely due to overestimating high, cold clouds and hydrometeors.

In summary, the individual valid-time brightness temperature plots reinforce the conclusions drawn from the composite images. Both HFSA and HFSB models capture the overall storm structures and temperature patterns reasonably well but tend to overestimate the coldness. These differences are more pronounced for the HFSB model.

4. Target diagram

The target diagram is a graphical tool for assessing model performance by comparing a model's output (m) to reference data (r). It displays three statistical metrics: bias (B), unbiased root-mean-square difference (RMSD'), and total root-mean-square difference (RMSD). The diagram uses a Cartesian coordinate system, with the x-axis representing RMSD' and the y-axis representing B. The distance from any point to the origin equals the RMSD, which is related to B and RMSD' by the equation $\mathrm{RMSD}^2 = B^2 + \mathrm{RMSD}'^2$. The target diagram enables the comparison of multiple models, with the best-performing model being closest to the origin.

The provided Hurricane Lee target diagrams (Figure 6) compare the predictive performance of models HFSA and HFSB over forecast lengths of ALL, 24, 48, 60, and 72 hours. Both models had a negative bias, indicating overestimating the coldness. HFSA (red crosses) consistently outperforms HFSB (blue circles) in terms of lower total RMSD across all intervals, suggesting it has superior predictive skill. The HFSB model has a larger negative bias and a larger variability, compared to HFSA. These diagrams visually and quantitatively demonstrate HFSA's overall better accuracy and lower error compared to HFSB. The same conclusion can be drawn for Idalia (Figure 7) and generally for Ophelia (Figure 8), with a notable exception of the 48-hour forecast, where HFSA and HFSB performed essentially the same.

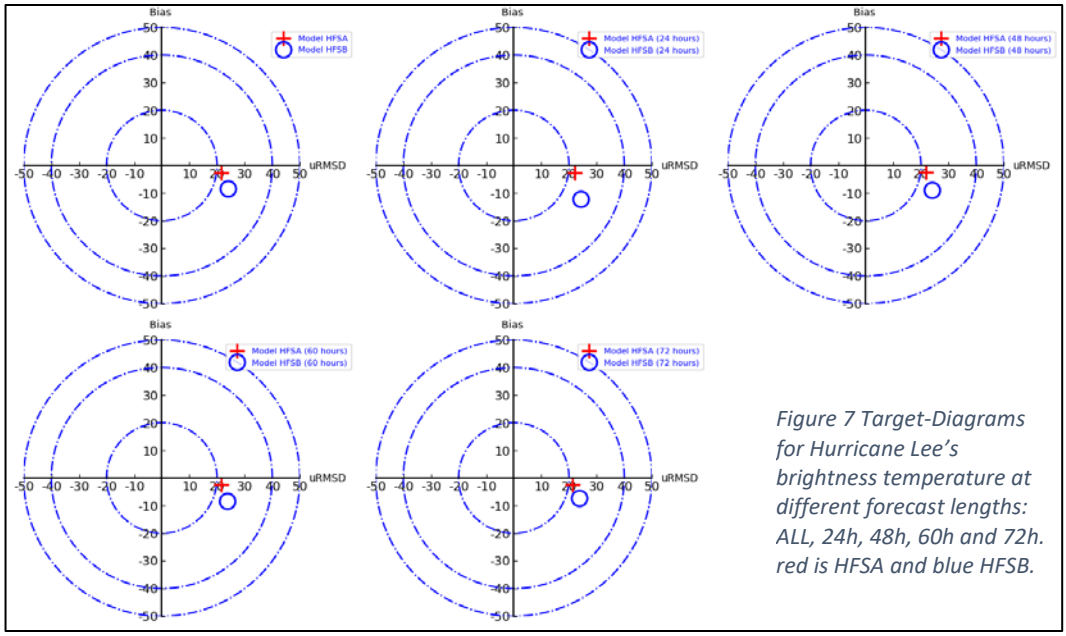

Figure 7 Target-Diagrams for Hurricane Lee's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB.

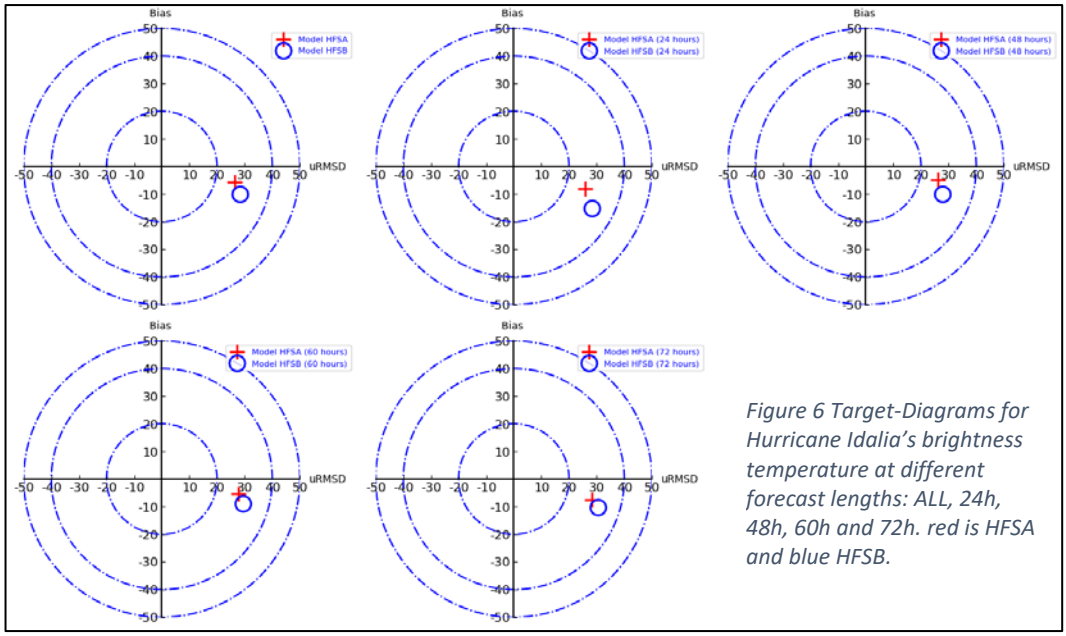

Figure 6 Target-Diagrams for Hurricane Idalia's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB.

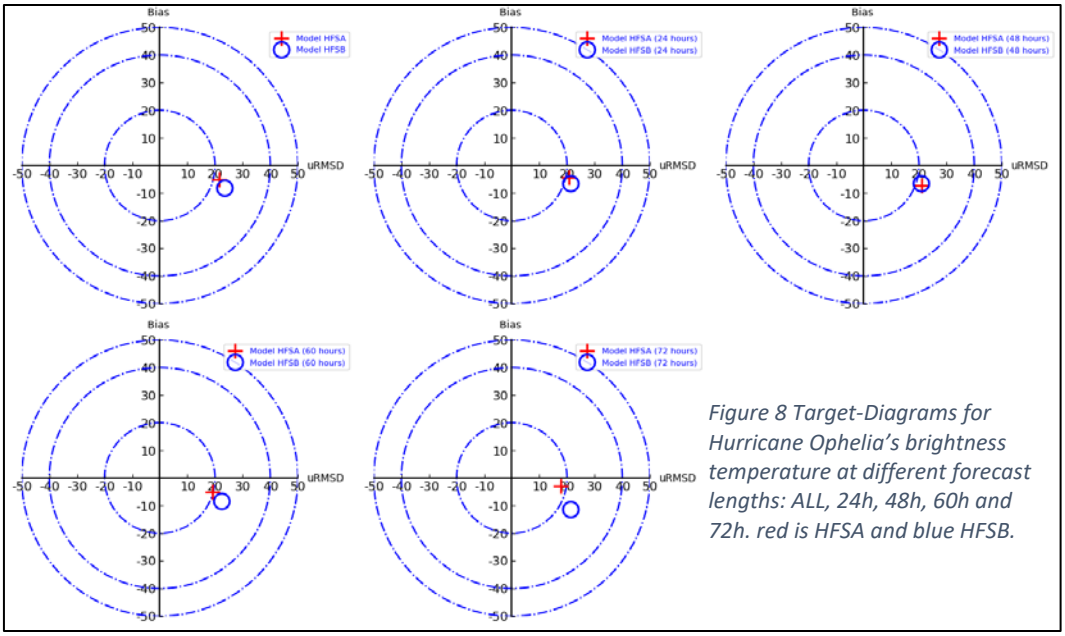

Figure 8 Target-Diagrams for Hurricane Ophelia's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB.

5. Taylor diagram

A Taylor diagram is a graphical tool used to evaluate the performance of different models by comparing their outputs to reference data using three key statistical metrics: the Pearson correlation coefficient (R), standard deviation (σ), and centered root-mean-square difference (RMSD'). The radial distance from the origin represents the model's standard deviation, the azimuthal angle indicates the correlation coefficient between the model and the observations, and the green contours represents RMSD. Closer proximity to the reference point (typically labeled "Observation") indicates better model performance, with a shorter distance signifying lower error and an angle closer to the reference line indicating higher

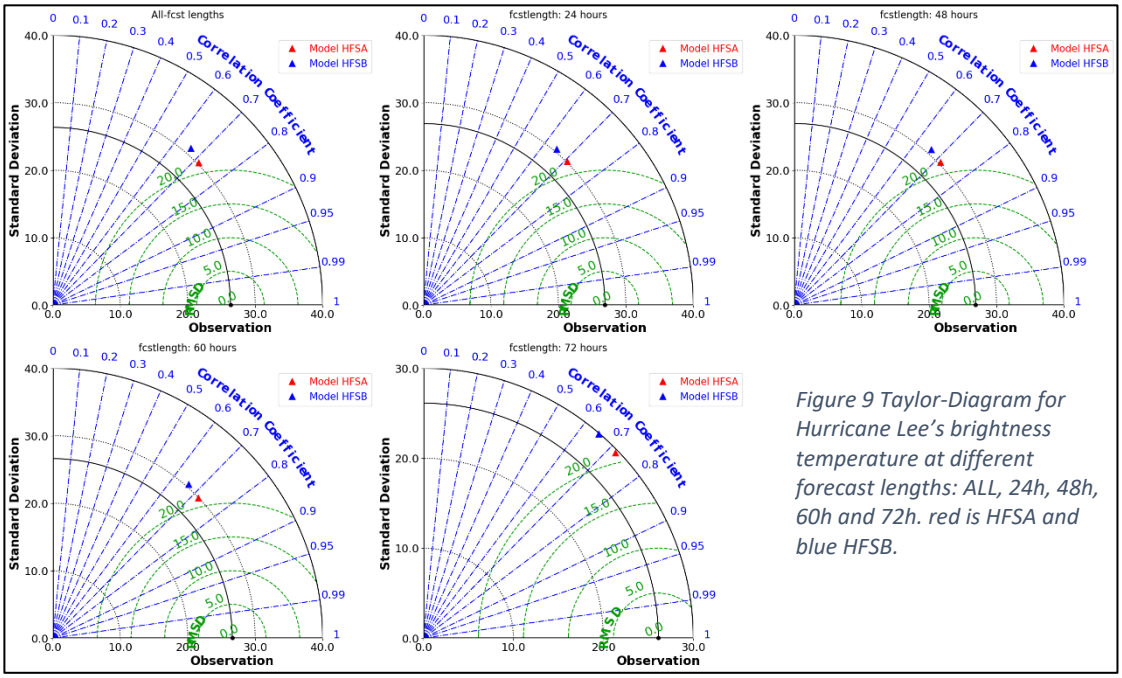

Figure 9 Taylor-Diagram for Hurricane Lee's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB.

correlation. Taylor diagrams offer a concise and intuitive way to compare the predictive skill of multiple models by visualizing their accuracy and variability relative to observed data.

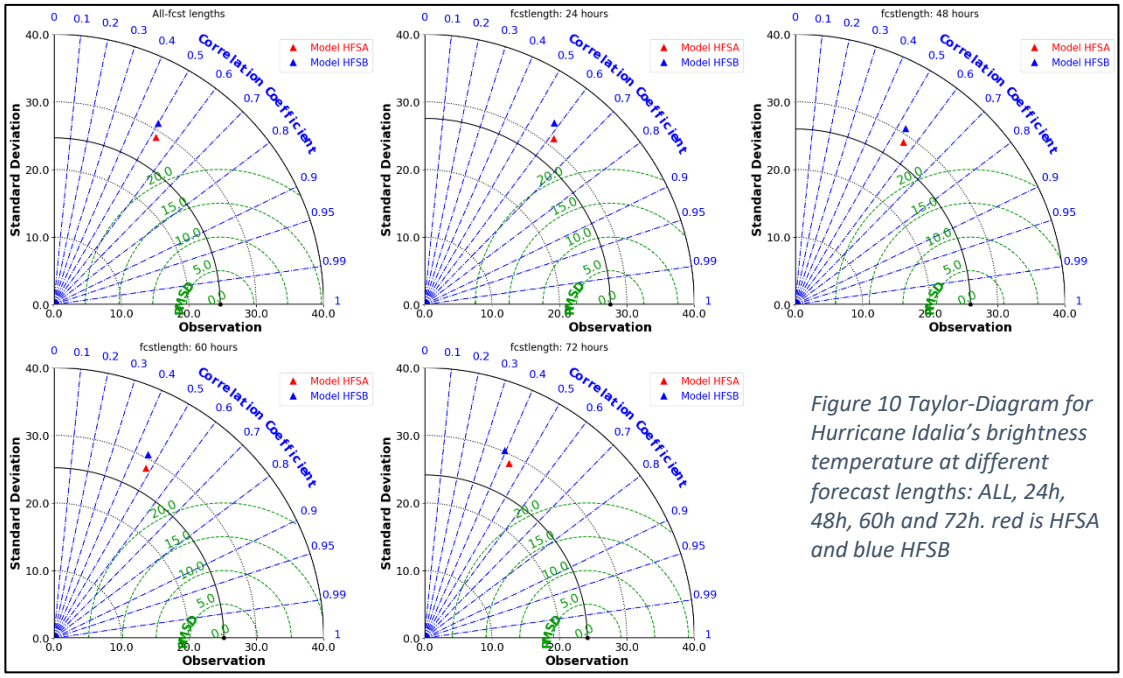

Figure 10 Taylor-Diagram for Hurricane Idalia's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB

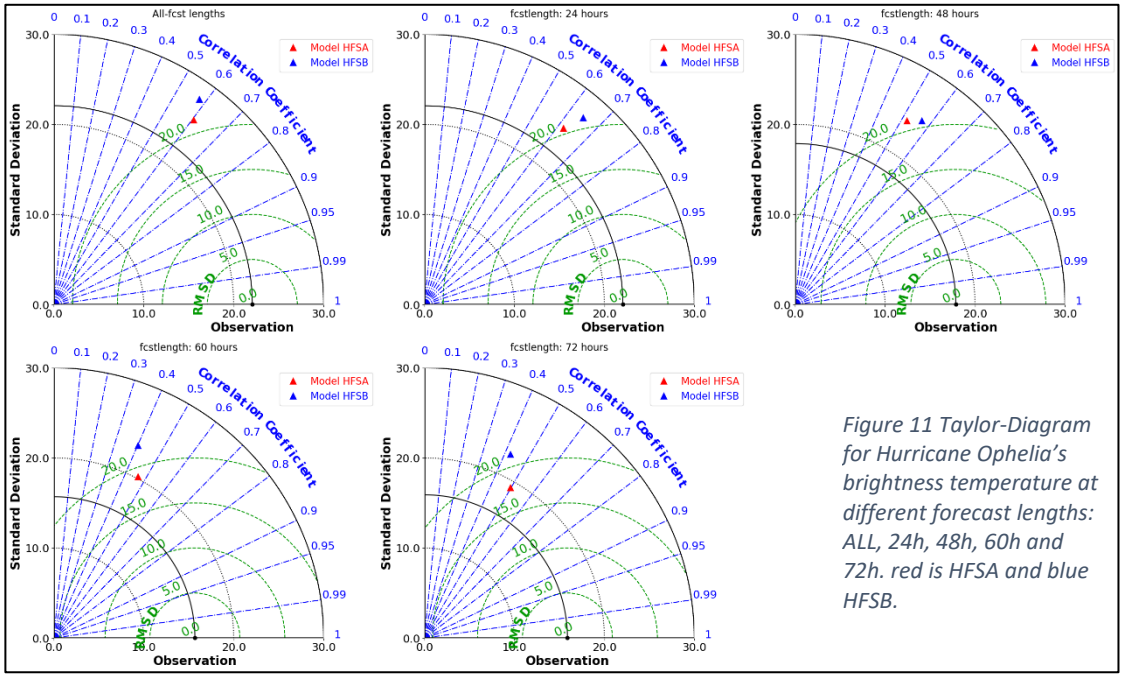

Figure 11 Taylor-Diagram for Hurricane Ophelia's brightness temperature at different forecast lengths: ALL, 24h, 48h, 60h and 72h. red is HFSA and blue HFSB.

The Taylor diagram (Figure 9) compares the predictive performance of models HFSA and HFSB over forecast lengths of 24, 48, 60, and 72 hours for Hurricane Lee. Across all intervals, HFSA (red triangles) consistently outperforms HFSB (blue triangles). HFSA shows higher correlation coefficients, indicated by

its closer proximity to the reference point (observations) along the x-axis reflecting more accurate variability in predictions. The green contours further highlight HFSA's lower RMSD', demonstrating its superior error performance. Specifically, all-time and at 24, 48, 60, and 72 hours, HFSA remains closer to the origin, indicating better alignment with observed data. These consistent trends across all forecast lengths underscore HFSA's robust predictive skill and reliability over HFSB.

The Taylor diagrams for Hurricanes Idalia (Figure 10) and Ophelia (Figure 11) show similar conclusions to those of Hurricane Lee. For both storms, HFSA consistently demonstrates superior performance compared to HFSB across all forecast lengths. HFSA maintains higher correlation coefficients and lower standard deviations, indicating more accurate and reliable predictions. The consistent trends across different storms reinforce HFSA's robustness and reliability as a predictive model, effectively summarized by the Taylor diagrams' visual representation of superior accuracy and consistency in model predictions compared to HFSB.

6. FSS

The Fractions Skill Score (FSS) is a spatial verification metric developed to evaluate the performance of high-resolution forecasts of precipitation or other spatially continuous fields. Traditional point-to-point verification metrics often struggle with the "double penalty" issue when applied to high-resolution forecasts, where small spatial displacements are harshly penalized. The FSS addresses this by comparing the forecast and observed fields over a neighborhood or region, rather than at individual grid points.

The FSS compares the forecast and observed fractional coverages of an event (e.g., brightness temperature exceeding a threshold) within successively larger spatial scales or neighborhoods. It is calculated as:

$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}{\frac{1}{N}\sum_{i=1}^{N}F_i^2 + \frac{1}{N}\sum_{i=1}^{N}O_i^2}$$

Where N is the number of neighborhoods, F_i is the forecast fraction, and O_i is the observed fraction for neighborhood i. The numerator represents the mean squared difference between forecast and observed fractions. The denominator is the largest possible value this could take.

The FSS allows users to identify the spatial scales at which a forecast exhibits useful skill by examining the variation of FSS with neighborhood size. An FSS close to 1 at small scales indicates the forecast captured small-scale features well.

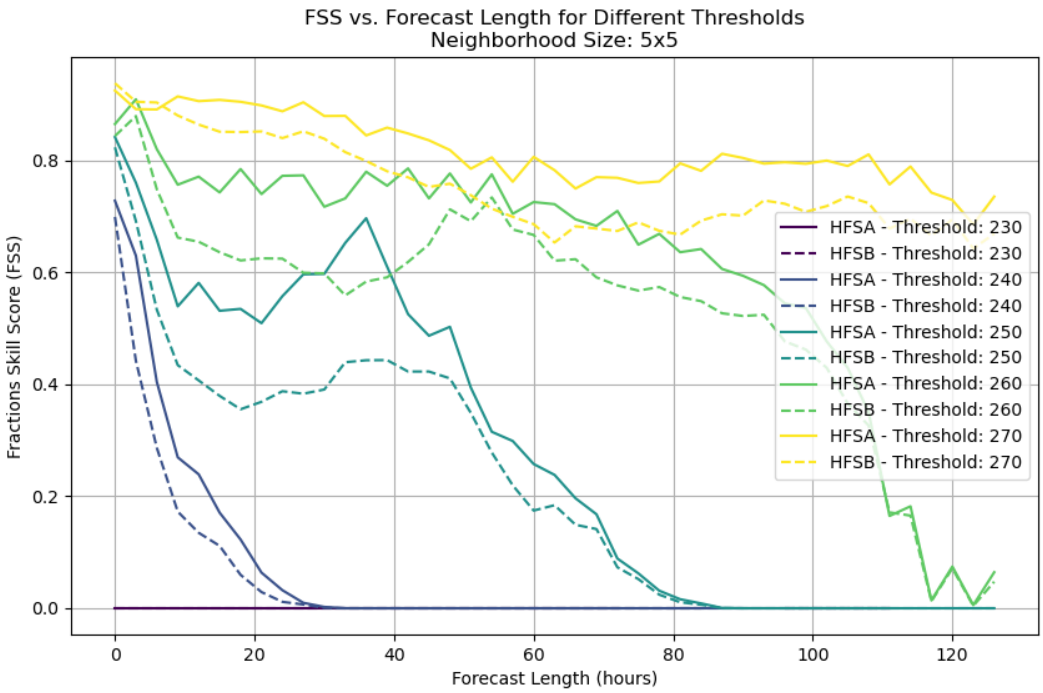

*Figure 12 Fractions Skill Score (FSS) vs. Forecast Length for Different Thresholds for Hurricane **Idalia** brightness temperature.*

The Fractions Skill Scores (FSS) shown in Fig. 12 illustrate the accuracy of two forecasting methods, HFSA (solid lines) and HFSB (dashed lines), over varying forecast lengths and thresholds for Hurricane Idalia. Initially, both methods exhibit high skill, but HFSA consistently demonstrates superior performance with higher FSS values across all thresholds. As forecast length increases, FSS generally declines for both methods, indicating reduced skill. This decline is more pronounced in HFSB, which shows lower initial FSS values and a steeper drop-off compared to HFSA. However, there are notable exceptions to this trend: HFSA at the 250 K threshold shows a modest increase in FSS between 20 and 35 hours, and HFSB at the 260 K threshold exhibits a similar rise between 35 and 55 hours. These temporary improvements may reflect periods when storm structure became more organized or when synoptic forcing improved model coherence. They might also indicate transitions out of model spin-up phases or temporary alignment with observed mesoscale cloud evolution. However, it is hard to diagnose the exact cause without extensive analysis of the storm environment, model dynamics, and observational data during those periods.

The performance difference is notable across all thresholds, particularly at higher ones (260 and 270), where HFSA maintains significantly better skill over longer periods. For the 230 K threshold, both methods exhibit no skill, with FSS values near zero across all forecast lead times. At 240 K, skill declines rapidly, approaching zero around 20 hours, though HFSA still outperforms HFSB. At the 250 K threshold, both methods maintain better performance, with FSS values remaining generally above 0.4 up to 50 hours, again with HFSA showing consistently higher skill. For thresholds 260 and 270, the best performance is observed, with HFSA maintaining FSS values above 0.8 initially and remaining relatively stable over time, whereas HFSB, while performing well, lags behind HFSA. Overall, HFSA exhibits more stable performance, particularly for forecast lengths beyond 80 hours, highlighting its superior long-term reliability. In contrast, HFSB shows more variability and a more significant drop in skill, especially at higher thresholds. This analysis underscores HFSA's reliability over HFSB, demonstrating better skill retention and higher accuracy

across various thresholds and forecast lengths. Therefore, for longer-term forecasts, particularly with higher thresholds, HFSA proves to be the more dependable method, consistently delivering better performance and maintaining higher skill levels over extended periods. This suggests that HFSA is more effective in capturing the spatial distribution of forecasted events, making it the preferred choice for accurate weather forecasting.

The FSS figure for Hurricane Lee (Figure 13) corroborates the earlier findings, revealing a similar pattern of HFSA's superiority over HFSB. Initially, both methods exhibit high skill, but HFSA consistently shows higher FSS values across all thresholds, indicating greater initial accuracy. As forecast length increases, FSS

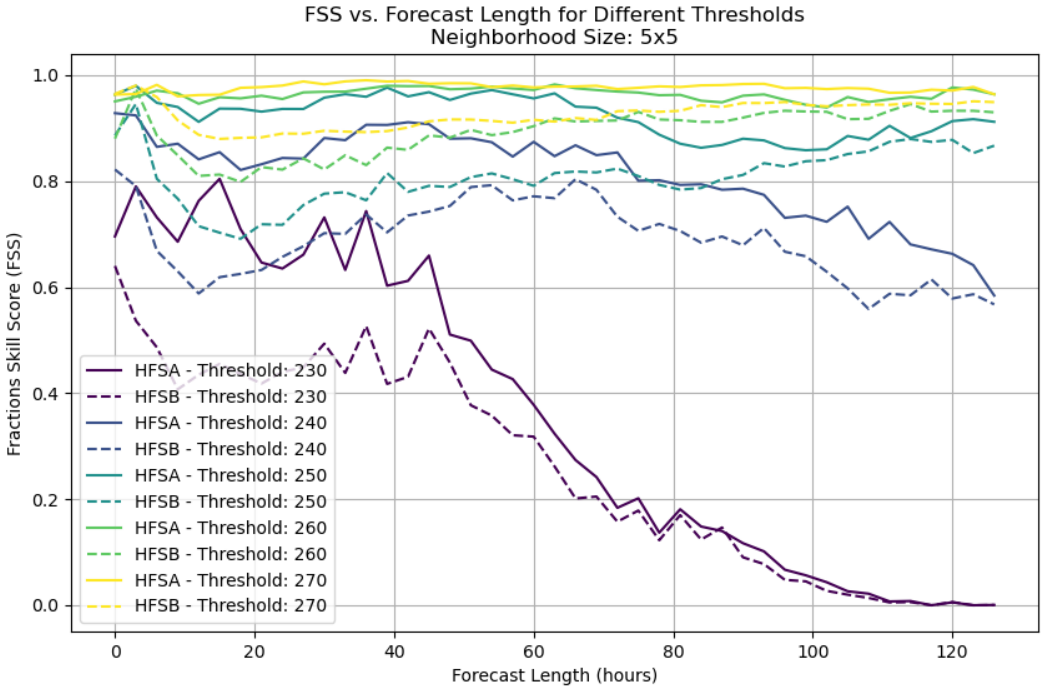

Figure 13 Fractions Skill Score (FSS) vs. Forecast Length for Different Thresholds for Hurricane Lee's brightness temperature.

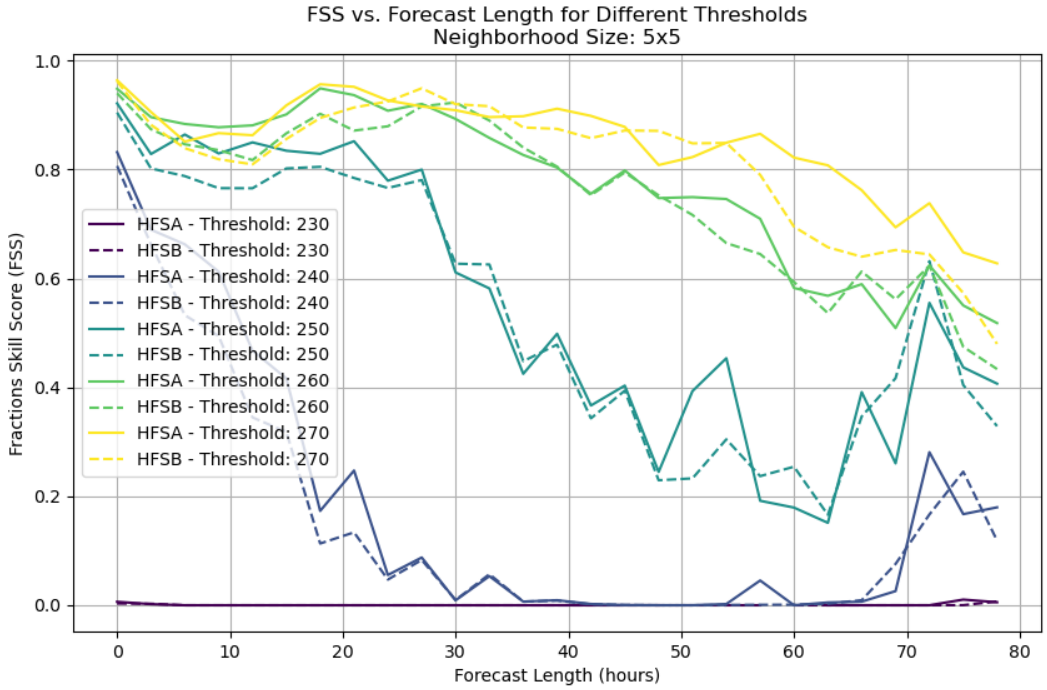

Figure 14 Fractions Skill Score (FSS) vs. Forecast Length for Different Thresholds for Hurricane Ophelia's brightness temperature.

values for both methods have a generally declining trend, but HFSA's decline is less steep compared to

HFSB. For threshold 230, both methods see a gradual drop in FSS, nearing zero by 120 hours, with HFSA maintaining higher scores. At threshold 240, HFSA continues to outperform HFSB, and for threshold 250, both methods sustain better skill, with HFSA leading. Higher thresholds (260 and 270) again highlight HFSA's advantage, maintaining FSS values close to or above 0.8 initially and showing stable performance over longer periods. HFSB, while performing well at these thresholds, consistently trails HFSA. For forecast periods beyond 80 hours, HFSA demonstrates greater stability and less variability, particularly at higher thresholds.

The previous FSS analyses showed HFSA consistently outperforming HFSB. However, the FSS figure for Ophelia (Figure 14) reveals a mixed performance. Initially, HFSA generally has higher FSS values, but as forecast length increases, the results vary by threshold. For the threshold of 240, HFSA maintains a slight edge before 60 hours. At thresholds above 250, both methods perform similarly. This mixed performance may be due to Ophelia's shorter duration — and hence a smaller sample size — as well as its weaker intensity compared to the other storms, resulting in a less organized vortex and a less conclusive FSS analysis.

7. Equitable Threat Score

The Equitable Threat Score (ETS), also known as the Gilbert Skill Score (GSS), is a widely used verification metric in the field of meteorology and climatology for assessing the skill of binary forecasts. It provides a means to quantify the performance of a forecast system by comparing the number of correctly predicted events (hits) against the number of expected hits due to random chance. The ETS is particularly useful when evaluating the accuracy of precipitation forecasts, as it considers the climatological frequency of the event being predicted. The score

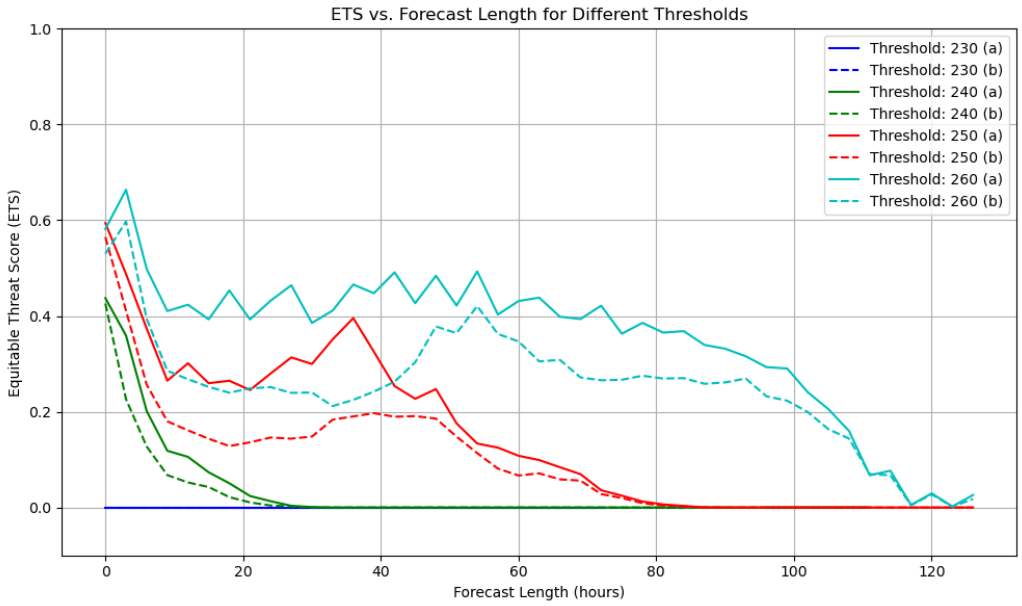

Figure 13 Equitable Threat Score vs. Forecast Length for Different Thresholds for Hurricane Idalia's brightness temperature.

ranges from -1/3 to 1, with 0 indicating no skill and 1 representing a perfect forecast.

Figure 15 shows Hurricane Idalia's ETS plotted against the forecast length (in hours) for different precipitation thresholds. Each color represents a specific threshold value, ranging from 230 to 260 in increments of 10. For each threshold, there are two lines: a solid line representing the HFSA model and a dashed line representing the HFSB model. As the forecast length increases, the ETS generally decreases for all thresholds and both models, indicating that forecast skill deteriorates with increasing lead time. Similar to Figure 12, there are exceptions to this trend—for instance, the ETS shows a modest increase for the HFSA threshold of 250K between 20–35 hours and for the HFSB threshold of 260K between 30–50 hours. Furthermore, both models exhibit near-zero forecast skill for brightness temperatures below 230K. The ETS values are higher for higher thresholds (e.g., 260) and decrease as the threshold decreases (e.g., 230). This suggests that both models perform better at predicting higher infrared brightness temperatures compared to lower temperatures.

The HFSA model (solid lines) generally outperforms the HFSB model (dashed lines) for most thresholds and forecast lengths, as indicated by the higher ETS values. The difference in performance between the two models appears to be more pronounced at shorter forecast lengths (up to around 40 hours) and for higher thresholds. At longer forecast lengths (beyond 80 hours), the difference in performance between the two models diminishes, and their ETS values converge for most thresholds. The ETS values for both models and all thresholds are relatively low (below 0.6), suggesting that there is significant room for improvement in the infrared brightness temperature forecasting skill of these models.

In summary, the HFSA model demonstrates better overall performance compared to the HFSB model as measured by ETS, particularly for higher infrared brightness temperature thresholds

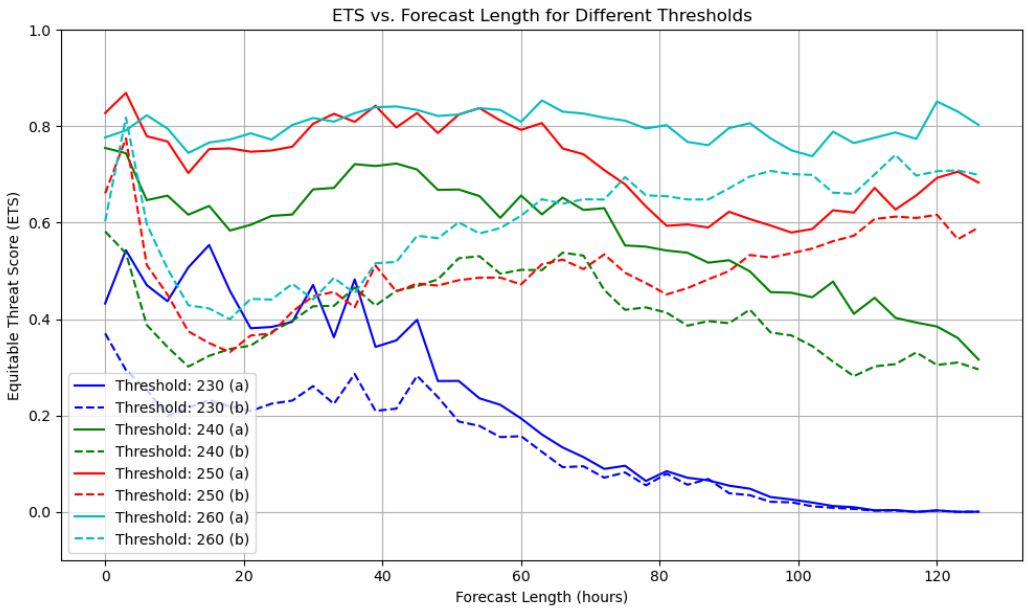

Figure 14 Equitable Threat Score vs. Forecast Length for Different Thresholds for Hurricane Lee's brightness temperature.

and shorter forecast lengths. However, both models struggle to maintain high ETS values at longer forecast lengths and lower thresholds.

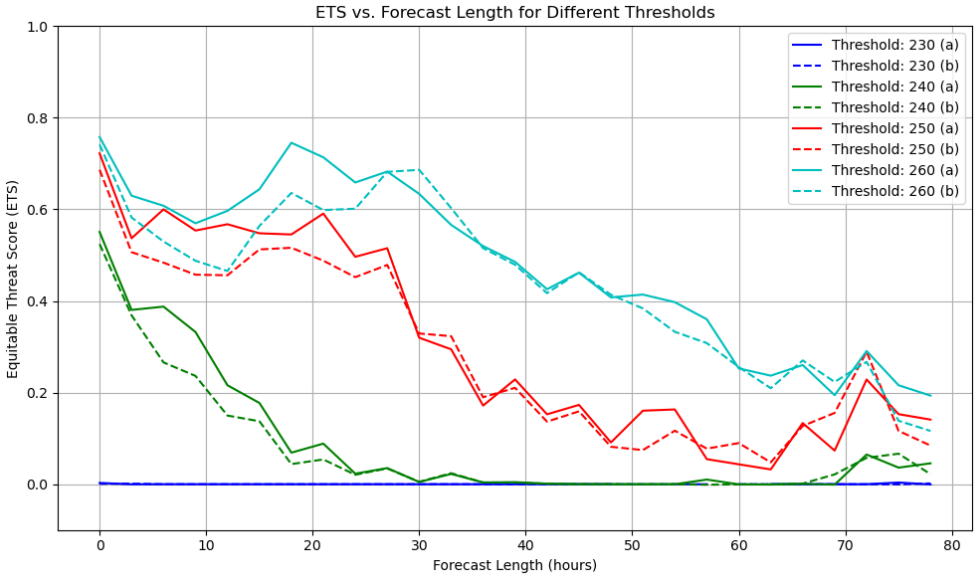

Figure 15 Equitable Threat Score vs. Forecast Length for Different Thresholds for Hurricane Ophelia's brightness temperature.

Hurricane Lee (Figure 16) exhibits higher ETS values than Hurricane Idalia, across all thresholds and forecast lead times for both models, indicating that the forecasts for Lee were generally more skillful. The HFSA model continues to outperform the HFSB model, particularly at shorter lead times and higher thresholds. However, the performance gap between the two models appears to be less pronounced for Hurricane Lee than for Idalia.

Hurricane Ophelia ETS (Figure 17) were generally lower than those in the Lee plot (Figure 16) but higher than those in the Idalia plot (Figure 15). The HFSA model tended to outperform the HFSB model across most thresholds and forecast lengths, although the difference in their performance was less pronounced and there are forecast times and thresholds where HFSB slightly exceeded HFSA.

Conclusion

Based on the comprehensive analysis of the HFSA and HFSB models' performance in predicting infrared brightness temperatures for Hurricanes Idalia, Lee, and Ophelia, several key conclusions can be drawn. The statistical metrics employed, including the Probability Density Function (PDF), composite images, target diagrams, Taylor diagrams, Fractions Skill Score (FSS), and Equitable Threat Score (ETS), generally demonstrate that the HFSA model outperforms the HFSB model across various thresholds and forecast lengths.

The PDF analysis revealed that both models tend to overestimate the cloud coverage and height, leading to larger cold temperature regions compared to the observed GOES-R satellite data. However, this discrepancy was more pronounced in the HFSB model than in the HFSA model. The composite images further corroborated these findings, with the HFSB plots appearing noticeably darker than the observed data, indicating a systematic bias towards overestimating the coldness of the brightness temperatures and, consequently, the extent of high, cold clouds and hydrometeors.

The target diagrams provided a quantitative assessment of the models' performance, with the HFSA model consistently exhibiting lower total RMSD values across all forecast intervals, suggesting superior predictive skill. The HFSB model, on the other hand, showed larger negative biases and greater variability compared to the HFSA model. Similarly, the Taylor diagrams demonstrated that the HFSA model maintains higher correlation coefficients and lower standard deviations, indicating more accurate and reliable predictions.

The FSS analysis further highlighted the HFSA model's superior performance. The HFSA model maintained higher skill levels over extended periods, showing less pronounced skill degradation over time compared to the HFSB model. This trend was consistent across Hurricanes Idalia and Lee, although the analysis of Hurricane Ophelia revealed a mixed performance, possibly due to the storm's shorter duration and weaker intensity.

Finally, the ETS analysis confirmed the HFSA model's better overall performance compared to the HFSB model, particularly for higher infrared brightness temperature thresholds and shorter forecast lengths. However, both models struggled to maintain high ETS values at longer forecast lengths and lower thresholds, indicating room for improvement in their forecasting skill.

In conclusion, the comprehensive evaluation of the HFSA and HFSB models' performance in predicting infrared brightness temperatures for three Atlantic hurricanes clearly demonstrates the superiority of the HFSA model. The HFSA model consistently outperforms the HFSB model across various statistical metrics, thresholds, and forecast lengths, exhibiting better predictive skill, lower biases, and higher reliability. However, both models show a tendency to overestimate cloud coverage and height, with this discrepancy being more evident in the HFSB model. Despite the HFSA model's better performance, there remains significant room for improvement in the forecasting skill of both models, particularly at longer forecast lengths and lower thresholds. Future research should focus on refining the models' parameterizations and algorithms to address these limitations and enhance their overall predictive capabilities. Additionally, extending the analysis to a larger sample of storms and exploring the models' performance in different basins and under varying environmental conditions would provide a more comprehensive assessment of their strengths and weaknesses, ultimately contributing to the development of more accurate and reliable hurricane forecasting tools

Discussions

1. Uncertainties

While this study provides valuable insights into the performance of the HFSA and HFSB models in predicting infrared brightness temperatures for three Atlantic hurricanes, it is crucial to acknowledge and discuss the uncertainties associated with the analysis and results.

An additional source of uncertainty arises from differences between the HFSA and HFSB configurations beyond the microphysics schemes. While our primary focus was on the impact of GFDL single-moment (HFSA) versus Thompson double-moment (HFSB) microphysics, other distinctions—such as the boundary layer parameterization settings (e.g., `tc_pbl=1` in HFSB) and tuning for entrainment in HFSA—may have contributed to the observed differences. Furthermore, both models were run with cycled data assimilation, which could influence the initial conditions and model performance. These factors introduce complexities in attributing differences solely to microphysics, and future studies should explore the full configuration space to better isolate individual contributions. Furthermore, uncertainty can arise from the compatibility between the microphysics schemes used in the models and the assumptions within the Community Radiative Transfer Model (CRTM). HFSA and HFSB use the GFDL and Thompson microphysics schemes, respectively, which differ in their representation of hydrometeors. While CRTM provides tailored optical property tables for each scheme, mismatches in particle size distributions, densities, and phase assumptions may still occur. These inconsistencies can affect the accuracy of the simulated brightness temperatures and lead to biases in model–observation comparisons, representing an inherent limitation in satellite-based model verification given the current state of the art.

Uncertainty also stems from the limited sample size of hurricanes considered in the study. The evaluation focused on three specific storms: Hurricanes Idalia, Lee, and Ophelia. Although these storms provided a diverse set of conditions for assessing the models' performance, they may not be fully representative of the wide range of hurricane characteristics and behaviors observed in the Atlantic basin. Extending the analysis to a larger sample of storms across multiple seasons would help to reduce this uncertainty and provide a more robust assessment of the models' capabilities.

Another source of uncertainty arises from the inherent limitations of the satellite observations used as reference data. The GOES-R satellite imagery, while providing high-resolution measurements of infrared brightness temperatures, may be subject to various sources of error, such as instrument noise, calibration uncertainties, and atmospheric attenuation. These errors can introduce discrepancies between the observed and simulated brightness temperatures, potentially affecting the evaluation of the models' performance.

The study also relies on a set of statistical metrics to quantify the models' performance, including the PDF, composite images, target diagrams, Taylor diagrams, FSS, and ETS. While these metrics provide valuable information about the models' skill and accuracy, they may not capture all aspects of the models' behavior and may be sensitive to the choice of thresholds and forecast lengths. The use of additional metrics or the exploration of alternative evaluation frameworks could help to reduce the uncertainty associated with the choice of performance measures.

Moreover, the study focuses on the evaluation of infrared brightness temperatures as a proxy for cloud coverage and height. While this approach provides valuable insights into the models' ability to simulate hurricane structure and intensity, it does not directly assess their performance in predicting other critical hurricane characteristics, such as wind speed, precipitation, and storm surge. Integrating additional variables and evaluation metrics could provide a more comprehensive assessment of the models' uncertainties and limitations.

Finally, the study does not explicitly address the potential impact of uncertainties in the initial conditions and boundary conditions used to drive the HFSA and HFSB models. Errors in the input data, such as atmospheric profiles, sea surface temperatures, and wind fields, can propagate through the models and

influence their predictive skill. Quantifying the sensitivity of the models' performance to these uncertainties would provide valuable information for guiding future model development and improvement efforts.

In conclusion, while this study offers important insights into the performance of the HFSA and HFSB models in predicting infrared brightness temperatures for Atlantic hurricanes, it is essential to recognize and consider the uncertainties associated with the analysis and results. Addressing these uncertainties through expanded storm samples, additional evaluation metrics, and sensitivity analyses will strengthen the robustness of the findings and contribute to the ongoing efforts to improve hurricane forecasting capabilities.

2. Connection to QPF T&E

The conclusions obtained in this study, which focused on evaluating the HFSA and HFSB models' performance in predicting infrared brightness temperatures, can indirectly contribute to improving quantitative precipitation forecasting (QPF) for hurricanes. While the study does not directly assess precipitation forecasts, the insights gained from analyzing the models' ability to simulate cloud coverage and height can inform efforts to enhance QPF accuracy.

One of the key findings of the study is that both the HFSA and HFSB models tend to overestimate cloud coverage and height compared to the observed GOES-R satellite data, with the HFSB model showing a more pronounced bias. This overestimation of cloud extent and coldness can have implications for QPF, as the presence and characteristics of clouds are closely linked to precipitation processes in hurricanes. By identifying and quantifying these biases, the study highlights areas where the models' parameterizations and algorithms could be refined to improve their representation of cloud physics and microphysics, which in turn can lead to better QPF performance.

Moreover, the study demonstrates that the HFSA model consistently outperforms the HFSB model in predicting infrared brightness temperatures across various thresholds and forecast lengths. This finding suggests that the HFSA model may have a more accurate representation of the atmospheric processes governing cloud formation and evolution in hurricanes. By extension, the HFSA model's superior performance in simulating cloud characteristics could potentially translate to better QPF skill, as the accurate prediction of cloud structure and intensity is a crucial prerequisite for reliable precipitation forecasts.

The evaluation metrics employed in the study, such as the FSS and ETS, provide valuable information about the models' ability to capture the spatial distribution and accuracy of predicted features. While these metrics were applied to infrared brightness temperatures in this study, they can also be used to assess the skill of QPF forecasts. The insights gained from analyzing the models' performance using these metrics can guide efforts to improve the spatial and temporal accuracy of precipitation predictions in hurricanes.

Furthermore, the study highlights the challenges associated with maintaining high forecast skill at longer lead times and lower thresholds, which is a common issue in QPF as well. By identifying these limitations, the study underscores the need for continued research and development efforts to improve the models' ability to predict precipitation accurately over extended forecast periods and for a range of intensity thresholds.

In conclusion, while the current study focuses on evaluating the HFSA and HFSB models' performance in predicting infrared brightness temperatures, the conclusions drawn from this analysis can indirectly contribute to improving QPF for hurricanes. The insights gained into the models' biases, relative performance, and limitations in simulating cloud characteristics can inform targeted efforts to refine the models' parameterizations, algorithms, and QPF capabilities. By addressing the identified issues and leveraging the superior performance of the HFSA model, researchers and forecasters can work towards enhancing the accuracy and reliability of quantitative precipitation forecasts for hurricanes, ultimately supporting better decision-making and preparedness efforts in the face of these severe weather events.

Future plan

Sensitivity to Microphysics Schemes and Diagnostics Studies: While the contrasting GFDL and Thompson microphysics schemes were evaluated, exploring the use of additional microphysics parameterizations could identify an optimal configuration for hurricane cloud/precipitation prediction. Conducting targeted diagnostics studies is recommended to gain deeper insights into how the different microphysics assumptions influence processes like condensation, evaporation, hydrometeor evolution, and precipitation efficiencies. Process-oriented diagnostics, such as analyzing the models' representation of convective/stratiform rainfall partitioning, ice/liquid hydrometeor profiles, and latent heating structures, could pinpoint strengths and deficiencies in the schemes. Such investigations may reveal compensating errors and provide guidance for microphysics scheme development tailored to tropical cyclone conditions.

Acknowledgement

This study was supported by the Developmental Testbed Center (DTC) Visitor Program. The DTC plays a crucial role in facilitating the transition of research advances into operational weather forecasting, and their support has been instrumental in enabling this evaluation of the HAFS configurations. The collaboration between the research team and the DTC has fostered a productive environment for advancing our understanding of tropical cyclone forecasting and identifying areas for improvement in the HAFS model.

Data Share:

https://coastal54-my.sharepoint.com/:f:/g/personal/sbao_coastal_edu/Euh8FNGfusdHt_XnMtc4l7cBJMYtuATUqnMri2m7vgyKKw?e=Ti1m65

Reference

Bao, S. , Z. Zhang, E. Kalina, and B. Liu, 2022: The Use of Composite GOES-R Satellite Imagery to Evaluate a TC Intensity and Vortex Structure Forecast by an FV3GFS-Based Hurricane Forecast Model. Atmosphere, 13, 126.

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.

Green, M., B. Veenhuis, J. Nelson, and M. Erickson, 2022: Evaluation and Guidance Development of HAFS Version A QPF over the Caribbean and Surrounding Regions. CIRES Rendezvous.

Ko, M.-C., F. D. Marks, G. J. Alaka, and S. G. Gopalakrishnan, 2020: Evaluation of Hurricane Harvey (2017) Rainfall in Deterministic and Probabilistic HWRF Forecasts. Atmosphere, 11, 666,

Lonfat, M., R. Rogers, T. Marchok, and F. D. Marks Jr, 2007: A parametric model for predicting hurricane rainfall. Mon. Weather Rev., 135, 3086–3097.

Marchok, T., R. Rogers, and R. Tuleya, 2007: Validation schemes for tropical cyclone quantitative precipitation forecasts: Evaluation of operational models for US landfalling cases. Weather Forecast., 22, 726–746.

Zhang, Z., Zhang, X., Liu, B., Mehra, A., Tallapragada, V., Gopalakrishnan, S., & Marks, F. D. (2022). Toward Initial Operational Capability: Progresses, Challenges, and Issues in Developing and Improving Hurricane Analysis and Forecast System (HAFS). Presented at the Unifying Innovations in Forecasting Capabilities Workshop, July 2022. Accessible at https://epic.noaa.gov/wp-content/uploads/2022/07/1.-HAFS-Zhang.pdf