Structural Forecasting for Short-term Tropical Cyclone Intensity Guidance

Trey McNeely,[a,d] Pavel Khokhlov, Niccolò Dalmasso,[b] Kimberly M. Wood,[c] and Ann B. Lee[a]

[a] *Carnegie Mellon University Department of Statistics and Data Science*

[b] *J.P. Morgan AI Research*

[c] *Mississippi State University Department of Geosciences*

[d] *Microsoft AI Development Acceleration Program*

Corresponding author: Ann B. Lee, annlee@stat.cmu.edu

ABSTRACT: Because geostationary satellite (Geo) imagery provides a high temporal resolution window into tropical cyclone (TC) behavior, we investigate the viability of its application to short-term probabilistic forecasts of TC convective structure to subsequently predict TC intensity. Here, we present a prototype model which is trained solely on two inputs: Geo infrared imagery leading up to the synoptic time of interest and intensity estimates up to 6 hours prior to that time. To estimate future TC structure, we compute cloud-top temperature radial profiles from infrared imagery and then simulate the evolution of an ensemble of those profiles over the subsequent 12 hours by applying a Deep Autoregressive Generative Model (PixelSNAIL). To forecast TC intensities at hours 6 and 12, we input *operational* intensity estimates up to the current time (0 h) and simulated future radial profiles up to +12 h into a "nowcasting" convolutional neural network. We limit our inputs to demonstrate the viability of our approach and to enable quantification of value added by the observed and simulated future radial profiles beyond operational intensity estimates alone. Our prototype model achieves a marginally higher error than the National Hurricane Center's official forecasts despite excluding environmental factors, such as vertical wind shear and sea surface temperature. We also demonstrate that it is possible to reasonably predict short-term evolution of TC convective structure via radial profiles from Geo infrared imagery, resulting in interpretable structural forecasts that may be valuable for TC operational guidance.

SIGNIFICANCE STATEMENT: This work presents a new method of short-term probabilistic forecasting for tropical cyclone (TC) convective structure and intensity using infrared geostationary satellite observations. Our prototype model's performance indicates that there is some value in observed and simulated future cloud-top temperature radial profiles for short-term intensity forecasting. The non-linear nature of machine learning tools can pose an interpretation challenge, but structural forecasts produced by our model can be directly evaluated and thus may offer helpful guidance to forecasters regarding short-term TC evolution. Since forecasters are time-limited in producing each advisory package despite a growing wealth of satellite observations, a tool that captures recent TC convective evolution and potential future changes may support their assessment of TC behavior in crafting their forecasts.

1. Introduction

Tropical cyclones (TCs) are powerful, organized systems that pose a major risk to coastal populations. Though many statistical models provide forecast guidance on future TC intensity change (e.g., the Statistical Hurricane Intensity Prediction Scheme [SHIPS]; DeMaria and Kaplan 1999), direct measurement of most predictors such as relative humidity or vertical wind shear used in such models is impossible due to the development of TCs over open ocean far from land-based observing networks (Gray 1979). Many predictors must be inferred through a combination of remote observation and dynamic models of ocean and atmospheric behavior.

Infrared (IR; 10.3-10.7μm) imagery from geostationary (Geo) satellites such as the Geostationary Operational Environmental Satellites (GOES) provides one of the few regular high-resolution observations of TC behavior over the open ocean with a historical record spanning decades (Knapp and Wilkins 2018; Janowiak et al. 2020). Furthermore, modern Geo IR platforms such as GOES-16 provide observations at even greater spatial and temporal resolution (Schmit et al. 2017). Since cloud-top temperature is related to cloud-top height, low IR temperatures tend to indicate higher cloud tops and thus stronger convection, and convective structures are known to be related to TC intensity (Dvorak 1975; Olander and Velden 2007).

In light of this growing record of satellite observations, a broad array of recent works have explored the wealth of information contained in the *spatio-temporal* structure of Geo IR imagery. The Dvorak technique and more recent Advanced Dvorak Technique (ADT) have long related Geo

IR imagery to TC intensity (Dvorak 1975; Olander and Velden 2007), and more recent work has leveraged neural networks to improve the nowcasting accuracy of the ADT (AI enhanced Dvorak Technique; Olander et al. 2021). Here, we define "nowcasting" as estimating the current TC intensity based on intensity estimates up to 6 h prior and IR features up to the current time (0 h). Spatial analyses of IR imagery have been leveraged to improve forecasts of TC eye formation, a process related to intensification (DeMaria 2015; Knaff and DeMaria 2017). The deviation angle variance (DAV) technique, a measure of convective organization in IR imagery, contains valuable information for short-term ($\leq$24 h) TC intensity guidance (Hu et al. 2020). The shape and evolution of Geo IR radial profiles is known to relate to intensity and intensity change respectively (Sanabia et al. 2014; McNeely et al. 2020). In this work, we utilize the *evolution over time* of radial profiles (see Figure 1) to jointly forecast short-term TC intensity and structure changes. We leverage deep auto-regressive (AR) generative models to construct *interpretable and high-resolution* structural probabilistic forecasts, which display entire functions rather than time series of thresholded quantities, such as pixel counts beneath a given temperature threshold.

Concurrent with the rise of high-resolution Geo IR imagery is the growing application of convolutional neural networks (CNNs), powerful tools for performing prediction tasks with images as input. Predicting TC intensity from Geo IR data is an obvious candidate application; indeed, there are dozens of such works in the machine learning literature applying CNNs to this problem, including Pradhan et al. (2017); Combinido et al. (2018); Lee et al. (2019); Tian et al. (2020); Wang et al. (2020); and Zhang et al. (2021). These models achieve reasonable forecast accuracy via the traditional machine learning framework with a CNN taking IR imagery as input to directly predict intensity by, e.g., minimizing the average squared-error loss on independent test data. Explainable AI approaches may then use methods such as layer-wise relevance propagation, saliency maps, and activation maps to better understand *how* the model produced its point estimate (McGovern et al. 2019; Ebert-Uphoff and Hilburn 2020). For an example of explainable CNN-based TC intensity forecasting in the meteorological literature, see Griffin et al. (2022).

Our proposed pipeline takes a different approach to explainability—one which remains compatible with the above tools for insight into the relationships leveraged by CNNs. Our approach (i) utilizes a dimensionality-reducing functional transformation of IR imagery prior to analysis, and (ii) provides 12-hour ensemble forecasts of TC convective structure in addition to TC intensity.

First, we extract *scientifically-motivated functional features*, reducing the dimension of the problem (from 2-D images over time to 1-D functions over time) in a directly interpretable summary, rather than directly relying on the CNN to extract salient features from (high-dimensional and low-sample size) raw Geo IR imagery. These rich summary functions are derived from the ORB suite: Organization (e.g., DAV as a function of radius), Radial structure (e.g., the radial profiles examined in this work), and Bulk morphology (e.g., pixel counts as a function of a temperature threshold). Temporal sequences of radial profiles are highly relevant to both intensity and intensity change (Sanabia et al. 2014; McNeely et al. 2020, 2022). Temporal changes in these sequences of profiles can be visualized via Hovmöller diagrams, which are more readily digestible by users than inferring temporal patterns from animations of satellite imagery.

Second, we provide a probabilistic *structural forecast*, a prediction of an ensemble of possible TC convective evolution, rather than directly predicting future intensity from past IR structure and TC intensity. Our novel approach to intensity guidance via Geo IR imagery results in interpretable intensity forecasts such as "our model predicts short-term intensification due to the potential emergence of an eye-eyewall structure in the next 12 hours". Though methods such as layer-wise relevance propagation can provide further insight into the CNN's use of structural forecasts, the IR structural forecasts themselves are the core of our proposed intensity guidance pipeline.

Figure 2 outlines Section 3 via a schematic diagram of the structural forecasting to intensity forecasting pipeline. There are three main subsections:

(a) *Structural trajectories via ORB.* First, we apply the ORB framework (McNeely et al. 2019, 2020) to observed IR imagery to create a "structural summary" (Figure 1) of the spatio-temporal evolution of the present and recent past TC structure.

(b) *Structural forecasting with a deep autoregressive generative model.* Next, we propagate the observed IR structure up to 12 hours forward in time via a deep pixel-autoregressive model, which stochastically simulates an ensemble of possible trajectories of IR radial profiles.

(c) *Forecasting TC intensity via convolutional neural networks.* Finally, we input the observed structure, the forecasted structure, and TC intensity up to 6 hours prior to the current time into a nowcasting model to estimate the current intensity; we choose CNNs because they are easy to train and commonly used for image data. By filling in the missing $t+6$ hour and $t+12$ hour

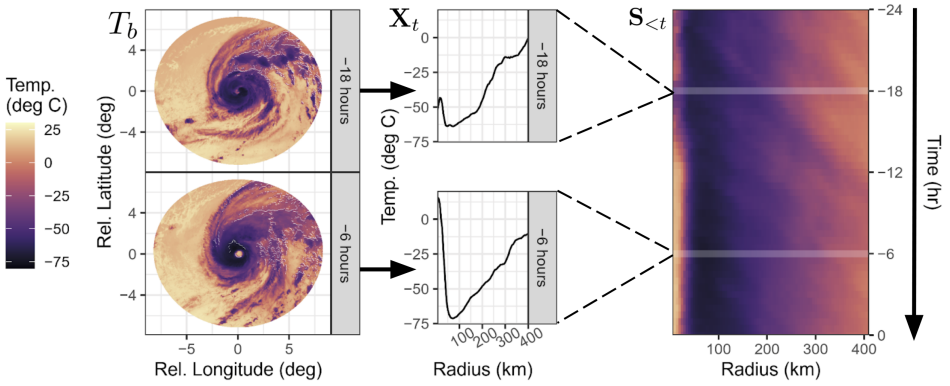


FIG. 1. **Evolution of TC IR Imagery as Structural Trajectories.** The raw data at each time t is a sequence of TC-centered cloud-top temperature images from GOES. We quantify the image at time t by its radial profile (X_t). The sequence of consecutive radial profiles, sampled every 2 hours, defines a *structural trajectory* or Hovmöller diagram ($S_{<t}$).

structure, we can then extend the nowcasting model from a nowcast for time t (i.e., hour 0) to a forecast at time $t+6$ hours and then to time $t+12$ hours.

Section 4 details the results of our prototype forecasting pipeline. The final Geo IR-based TC intensity guidance provides inherent measures of uncertainty and insight into the potential TC structural changes that influence a given forecast. The results in this work use proof-of-concept structural forecasting and a pipeline that relies solely on persistence predictors (i.e., prior intensity estimates) together with observed past and simulated future radial profiles; no environmental factors such as vertical wind shear or ocean heat content are included at this time. We demonstrate that a purely autoregressive prototype achieves a useful degree of forecasting accuracy.

2. Data

Our model relies on two data sources: sequences of Geo IR imagery captured by GOES satellites and past TC intensity. For training and verification (that is, model selection), we use NOAA's Hurricane Database 2 (HURDAT2; Landsea and Franklin 2013) because that database provides the post-season best estimates of TC intensities. For forecasting, we rely on *operational* TC intensity estimates, the CARQ entries from the Naval Research Laboratory's Automated Tropical Cyclone Forecast (ATCF) operational "A-deck" files (Sampson and Schrader 2000) to assess model performance under real-time conditions.

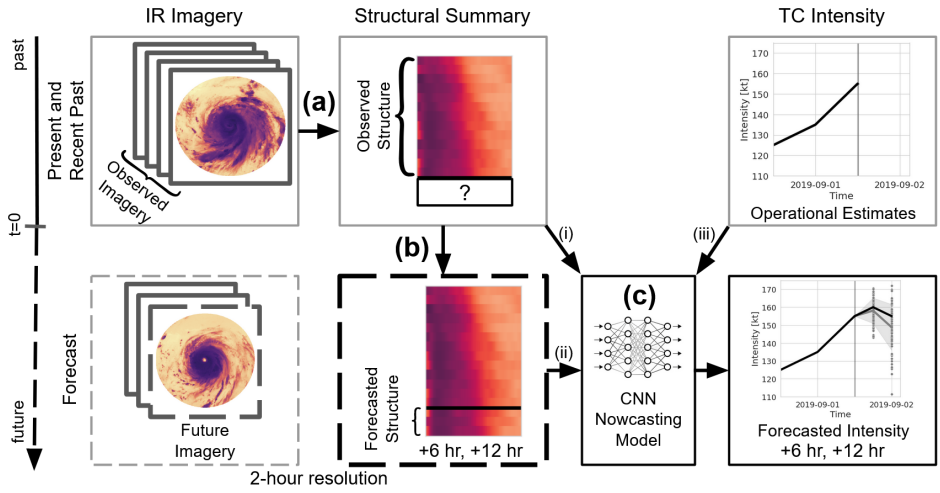


FIG. 2. **TC Intensity Guidance via Structural Forecasting.** Outline of the structural forecasting to intensity guidance pipeline. **(a)** ORB functions are used to quantify the evolution of spatio-temporal convective structure, linking IR imagery to Structural Summaries. **(b)** We generate structural forecasts by projecting the ORB functions into the future via a deep autoregressive model, thereby filling in the missing +6 hour and +12 hour structure. **(c)** A CNN nowcasting model then forecasts intensities at +6 to +12 hours from three sources of inputs: (i) observed structure, (ii) forecasted structure, and (iii) operational intensity estimates up through the current time.

GOES IR imagery is available through NOAA's Merged IR (MERGIR) database (Janowiak et al. 2020) at 30-minute×4-km resolution over the North Atlantic (NAL) basin from 2000–2020. For each TC, we download ~2,000 km×2,000 km "stamps" of IR imagery centered on the TC location at a 30-minute temporal resolution. Figure 1 (left) shows two such stamps after an 800-km radius mask is applied. For this work, we sample the 30-minute data at 2-hour resolution because of periodic corruption of the imagery in the MERGIR database (Liu 2021).

During training, we linearly interpolate TC location and intensity from HURDAT2 to obtain locations and intensities for non-synoptic times; however, model assessment is restricted to synoptic times. We include TC lifetimes between the first synoptic time at which intensity reaches at least 35 kt and the last synoptic time at which intensity is at least 35 kt; note that this can result in the inclusion of TCs < 35 kt if the TC decays and then re-intensifies.

Finally, we rely on NHC's official forecast verification to assess our model's performance. We also draw on the SHIPS developmental database's 200-850-hPa vertical wind shear values calculated within a 200-800-km annulus from the TC center as reference during model validation due to the known impact of shear on TC convective structure (DeMaria 2018).

3. Methods

As outlined in Figure 2, we first construct a summary of IR structural evolution (**a**; Section 3a). We then train a stochastic autoregressive model, which is an explicit likelihood model (of structural trajectories) that we can use to simulate probable IR structural evolution (**b**; Section 3b). Finally, we combine observed and forecasted structure with operational intensity estimates up to and including the current time to provide interpretable short-term intensity guidance, based solely on Geo IR imagery and operational intensity estimates (**c**; Section 3c).

a. Structural Trajectories via ORB

Operational forecasting of TC intensity is a human-in-the-loop process and thus places a premium on guidance interpretability. In this spirit, the ORB framework (Organization, Radial structure, Bulk morphology) summarizes 2-D imagery via continuous 1-D functions to enable static visualization of spatio-temporal patterns in TC development via Hovmöller diagrams (Hovmöller 1949). Our past work focused on the rich quantification of spatial information in Geo IR imagery (McNeely et al. 2019, 2020). More recently, we demonstrated the value of *temporal* patterns in ORB functions (McNeely et al. 2022), specifically the radial profile.

The radial profile of brightness temperature $\overline{BT}(r) = \frac{1}{2\pi}\int_0^{2\pi} T_b(r,\theta)d\theta$ captures the structure of cloud-top brightness temperature (T_b) as a function of radius r from the TC center and serves as an easily interpretable description of the depth and location of convection near the TC core (Sanabia et al. 2014; McNeely et al. 2020). The radial profiles are computed at 5-km resolution from 0-400 km ($d = 80$) (Figure 1, center); we denote the summary of convective structure at each time t by $\mathbf{BT}_t$. The *structural trajectory* is then defined as the 24-hour sequence of present and 12 preceding radial profiles at a 2-hour resolution:

$$\mathbf{S}_{<t} = (\mathbf{BT}_{t-24h}, \mathbf{BT}_{t-22h}, \ldots, \mathbf{BT}_t) \tag{1}$$

We visualize such a trajectory with a Hovmöller diagram (see Figure 1, right).

McNeely et al. (2022) demonstrated a relationship between TC intensity change and Hovmöller diagrams of radial profiles. However, the radial profile, if averaged over all angles, will disregard asymmetry within the original 2-D images, which can degrade performance for cases affected

by strong vertical wind shear. In this work, we instead compute a separate radial profile for each geographic quadrant (NE, NW, SE, SW) to capture asymmetries via the differences between quadrants. We use geographic quadrants instead of motion-relative or shear-relative quadrants because the directions of motion and shear are unstable when the magnitudes of those vectors are small.

b. Structural Forecasting via Deep Autoregressive Generative Model

The crucial step in our guidance framework is the propagation of radial profiles into the near future. The Hovmöller diagram captures the spatio-temporal evolution of the TC over an extended period of time; that is, we can summarize TC development by an easily-interpretable image. By treating the structural trajectory as an image, where the y-axis corresponds to the passage of time, forecasting radial profiles becomes equivalent to an image completion problem. That is, we predict the missing pixels at the bottom of an image (forecasted structure) given those at the top (observed structure). Image completion is an active research area in machine learning; here we focus on a state-of-the-art model in the class of pixel-autoregressive models (Van Oord et al. 2016).

Pixel-autoregressive models impose an ordering on the pixels of an image, such as a raster-scan ordering (left-to-right, top-to-bottom). Let each pixel in a four-quadrant radial profile trajectory be represented by $\mathbf{x}_i := (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, where i is the index in the raster scan. The pixel-AR approach factors the joint distribution of pixel values in the image as a product of conditionals,

$$p(\mathbf{x}_1, ..., \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i | \mathbf{x}_{i-1}, ..., \mathbf{x}_1), \tag{2}$$

where the probability of each pixel value is conditioned on all previous pixels in the raster scan. Then, to generate a new radial profile, one simulates repeatedly from $p(\mathbf{x}_i | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)$; due to the raster-scan ordering, the distribution of a given pixel is not influenced by elements further down the sequence, hence enforcing causality.

The challenge of how to estimate the conditional likelihoods $p(\mathbf{x}_i | \mathbf{x}_{i-1}, ..., \mathbf{x}_1)$ has given rise to many flavors of pixel-autoregressive models, including PixelRNN (Van Oord et al. 2016), PixelCNN (Van den Oord et al. 2016), PixelCNN++ (Salimans et al. 2017), and PixelSNAIL (Chen et al. 2018). This work utilizes the last model, PixelSNAIL. There are two main ingredients

in the model: (i) causal convolution and (ii) self-attention. Causal convolution utilizes the same convolutional feature extraction outlined in Section c but masks each convolution so that each element in the raster sequence only receives information from previously generated sequences (e.g., Figure 3). Purely convolutional models, however, are restricted to small neighborhoods of pixels, leading to only a finite receptive field (area of the source image involved in a given convolution), and thus struggle with long-range dependencies in the conditional $p(\mathbf{x}_i|\mathbf{x}_{i-1},...,\mathbf{x}_1)$. PixelSNAIL, on the other hand, features a self-attention mechanism that leads to unbounded receptive fields with pinpointed access to information far away in the sequence; see Chen et al. (2018) for details on the PixelSNAIL architecture.

To ensure that cloud-top temperatures remain bounded, we rescale the data to the range $\mathbf{X} \in (0,1)^4$ and work in the logit-transformed space, $\mathbf{Z} = \log(\mathbf{X}/(1-\mathbf{X}))$; while values of $\mathbf{Z}$ are unbounded, $\mathbf{X}$ remains bounded in $(0,1)^4$ and are then transformed back to the temperature range observed in the training data. We model the density $p(\mathbf{z}_i|\mathbf{z}_{i-1},...,\mathbf{z}_1)$ as 4 independent mixtures of logistic distributions, one for each quadrant. That is, for pixel i in quadrant q,

$$
\begin{aligned}
p(z_{iq}|\mathbf{z}_{i-1},\ldots,\mathbf{z}_1) &= \sum_{k=1}^{K} \pi_{qk} f(z_{iq};\mu_{qk},s_{qk}) && \text{(mixture distribution)},\\
\text{where } f(z_{iq};\mu_{qk},s_{qk}) &= \frac{g_{qk}(z)}{s_{qk}(1-g_{qk}(z))^2}, && \text{(logistic distribution)}\\
g_{qk}(z) &= \exp\big(-(z-\mu_{qk})/s_{qk}\big),\\
\text{and } \sum_{k=1}^{K} \pi_{qk} &= 1.
\end{aligned}
$$

Thus, with four quadrants and K mixture components, the distribution $p(\mathbf{z}_i|\mathbf{z}_{i-1},...,\mathbf{z}_1)$ has $4(3K-1)$ parameters: a mean μ, a scale s, and a mixture coefficient π for each quadrant, with the constraint that the mixture coefficients in each quadrant sum to one. With $K = 3$ mixture components, this results in 32 parameters total. Each draw from the distribution is transformed back to the bounded space via the relationship $\mathbf{X} = \frac{1}{1+\exp(-\mathbf{Z})}$, then rescaled to the range of values observed in the input radial profiles.

This autoregressive model enables stochastic simulation of structural trajectories based on Geo IR persistence. For a given synoptic time, we can simulate many trajectories from the observed

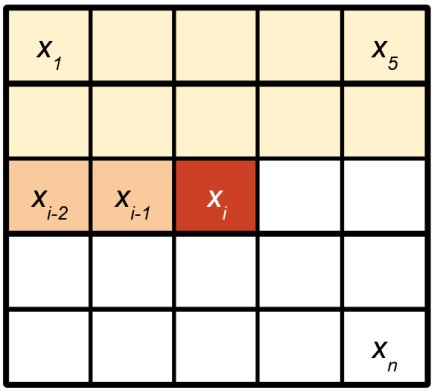


FIG. 3. **Masking in Pixel Autoregression.** Illustration of raster-scan ordering and the causal masking. Convolutions at index i only have access to pixel values in previous rows (earlier time points, color coded by yellow), and pixel values in the same row but to the left of pixel x_i (same time point, color coded by orange).

history and then feed each potential trajectory through the nowcasting model to obtain the associated intensity guidance. Via multiple simulations per forecast time, an ensemble forecast provides a measure of uncertainty in both structural trajectories and intensities while also offering insight into cases where the model over- or under-estimates intensity. For example, overestimates may be caused by too-low profile temperatures or overestimated symmetry between quadrants.

The structural forecasting model is trained on TCs from 2000-2012, with 2013-2020 withheld for testing. We train the model using input radial profiles calculated every 2 hours but test on synoptic times. Because AR models are likelihood-based, we can directly calculate and minimize the negative log-likelihood (NLL), a measure of the model's ability to generalize well on withheld data.

c. Nowcasting TC Intensity via Convolutional Neural Networks

Traditional linear models are attractive for reasons of interpretability and good performance in low sample size settings. However, linear models often struggle to capture the complex, time-varying processes which drive TCs. It is also unclear how to include the radial profile Hovmöller diagrams as inputs to a linear model without sacrificing interpretability. In this work, we instead consider a simple convolutional neural network to map observed IR trajectories ($\mathbf{S}_t$) to current intensities (Y_t). Because we treat time as a spatial dimension in these diagrams and a structural trajectory is represented as an image, a CNN will leverage both spatial and temporal patterns in the data.

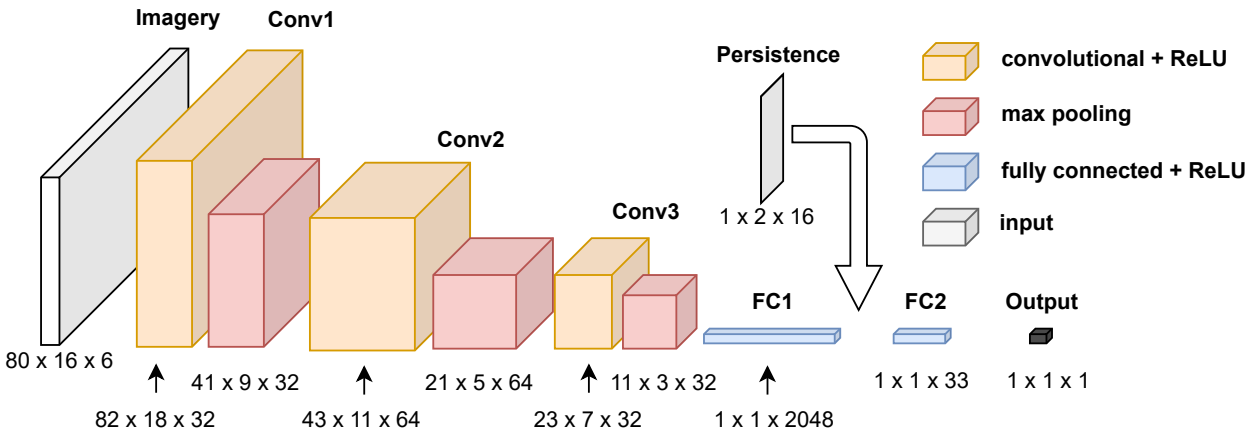


FIG. 4. **Nowcasting Model Architecture.** The convolutional neural network used for nowcasting consists of three convolution-max pool layer pairs (Conv#), fully connected layers (FC#), and a concatenation with persistence features between regressions.

CNNs operate by two main elements: convolutional layers and fully-connected layers. The convolutional layers first convolve each layer (here, each quadrant) with a library of filters (i.e., matrices whose entries are learned parameters); some of these filters may resemble familiar matrices, such as gradient approximators (e.g., Sobel matrices). After each convolutional layer, the image is pooled to reduce the image size and increase the receptive field of the next set of convolutions. In the final step, the results of all convolutions are passed into a fully-connected layer which approximates the relationship between the convolutional feature map and the response.

We augment the traditional CNN in two ways. First, we add two layers which encode the radial and temporal location of pixels within the image; regular CNNs are translation invariant, whereas patterns in TCs have different meaning depending on their location and time of occurrence (corresponding to column versus row index, respectively, in the Hovmöller diagram). Second, we augment the model output with relevant TC persistence features (intensities and intensity changes up to 6 hours prior to time t) before passing them to a second fully-connected layer. The final model is then

$$Y_t = f_{\text{nwcst}}(\mathbf{S}_{<t}, Y_{t-30h}, Y_{t-24h}, \ldots, Y_{t-6h}, \Delta Y_{t-30h}, \Delta Y_{t-24h}, \ldots, Y_{t-6h}) + \epsilon_t, \tag{3}$$

where $\mathbf{S}_{<t}$ is the 24-h structural trajectory up to current time (see Equation 1); the Y's are intensity data at 6-h resolution; the ΔY's are intensity changes interpolated (ip) from 6-h data to a 2-h resolution (e.g., $\Delta Y_{t-30h} = Y_{t-30h} - Y^{\text{ip}}_{t-32h}$), and ϵ is the prediction error.

Like the structural forecasting model, the nowcasting model is trained on TCs from 2000-2012, here by minimizing the mean squared error. The model is trained on data with a 2-hour resolution (rather than synoptic times alone) with intensities linearly interpolated to those times; we do not include non-synoptic times in the test TCs (2013-2020). The details of the CNN architecture are given in Figure 4.

d. From Nowcasting to Forecasting

Section 3c defines a nowcasting model for estimating intensity at time t (i.e., hour 0) by training on post-season (best-track or HURDAT2) intensities from -30 h to -6 h and imagery from -30 h to 0 h. After we have trained and validated the nowcasting model to estimate 0-h intensities, we apply the CNN nowcasting model to TC intensity forecasting. To forecast intensity at time $t+6$ h, we need the intensities at times $\leq t$ (in this work, we use operational intensities drawn from CARQ in the A-deck files when generating TC intensity *forecasts*) and structural trajectories at times $\leq t+6$ h (observed at times $\leq t$ and simulated at times from $t+2$ h to $t+6$ h). Using the structural forecasting model in Section 3b, we simulate many possible trajectories from times $t+2$ hr to $t+6$ hr. Each of these possible future trajectories is then passed to the nowcasting model to obtain a separate intensity forecast, giving an ensemble of possible intensities.

Our proposed framework for intensity forecasts at +6 and +12 h has two primary benefits: (i) by providing an additional *structural* forecast, we provide insight into potential TC evolution predicted by the model, such as deepening convection or the emergence of an eye; (ii) because the structural forecast is stochastic, we can straightforwardly assess the uncertainty in structural evolution over time and the associated uncertainty in the intensity forecasts.

4. Model Results

We first demonstrate the performance of our proposed model on specific cases (Hurricanes Jose [2017], Nicole [2016], and Dorian [2019]) in Section a, discussing both accuracy and the insight

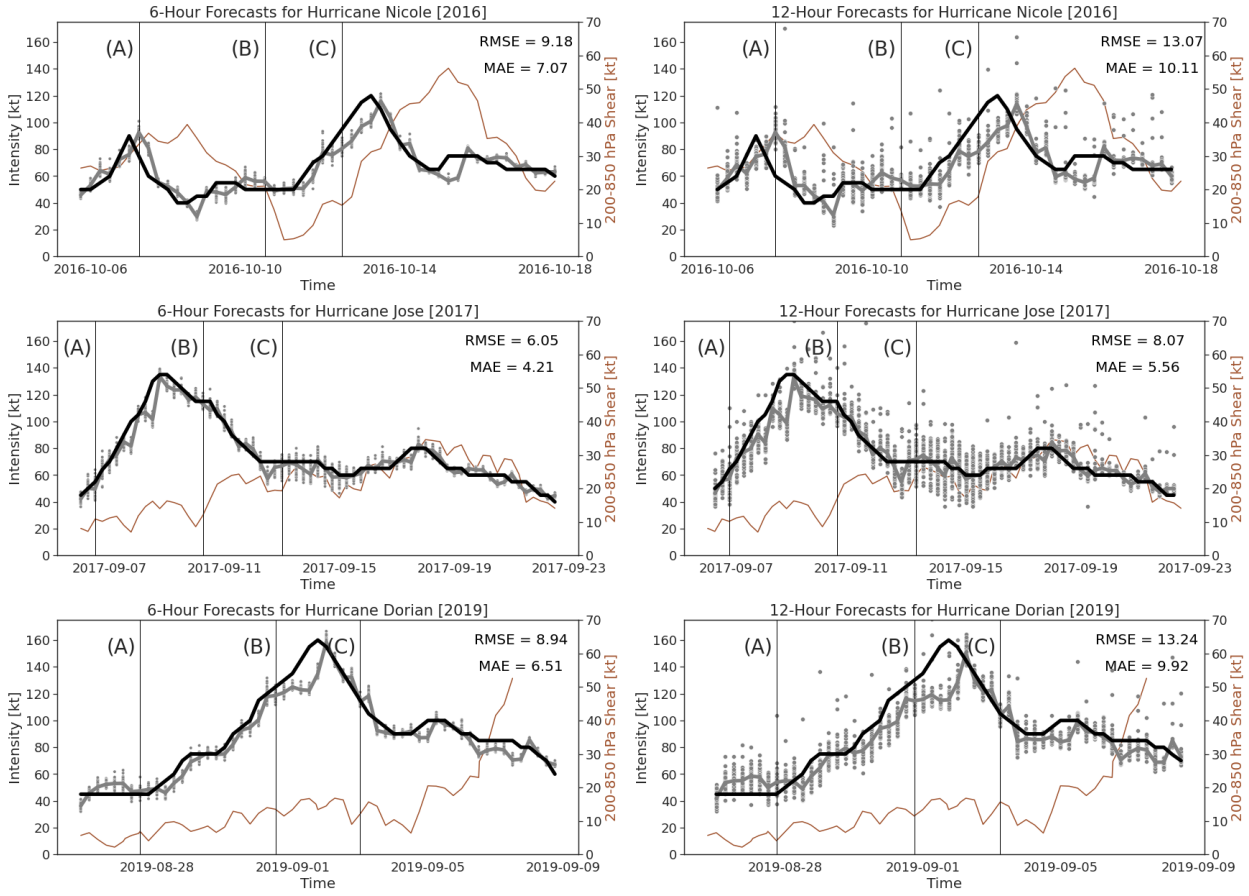


FIG. 5. **Case Studies:** Comparison of the 6-hour (left) and 12-hour (right) intensity forecasts (gray) from the structural forecasting model versus observed intensities (black). Vertical wind shear also shown (gold). The solid gray lines indicate the mean of 64 simulations at each time point. Reported errors indicate the error of the average model prediction, not individual simulations. The model captures the behavior of Hurricanes Jose [2017; center] and Dorian [2019; bottom] fairly well, but struggles with both rapid weakening events exhibited by Hurricane Nicole [2016; top] as well as low-intensity maintenance periods in Hurricanes Nicole and Dorian.

provided by structural forecasts at 6- and 12-hour lead times. We then assess the performance of the model during 2013-2020 in the North Atlantic basin at 6- and 12-hour lead times in Section b.

a. Case Studies

We examine Hurricane Jose (2017) due to the presence of high vertical wind shear which produces convective asymmetries not captured by the azimuthally-averaged radial profiles (i.e., not quadrant-based) of McNeely et al. (2022). Hurricane Nicole (2016) was selected due to undergoing

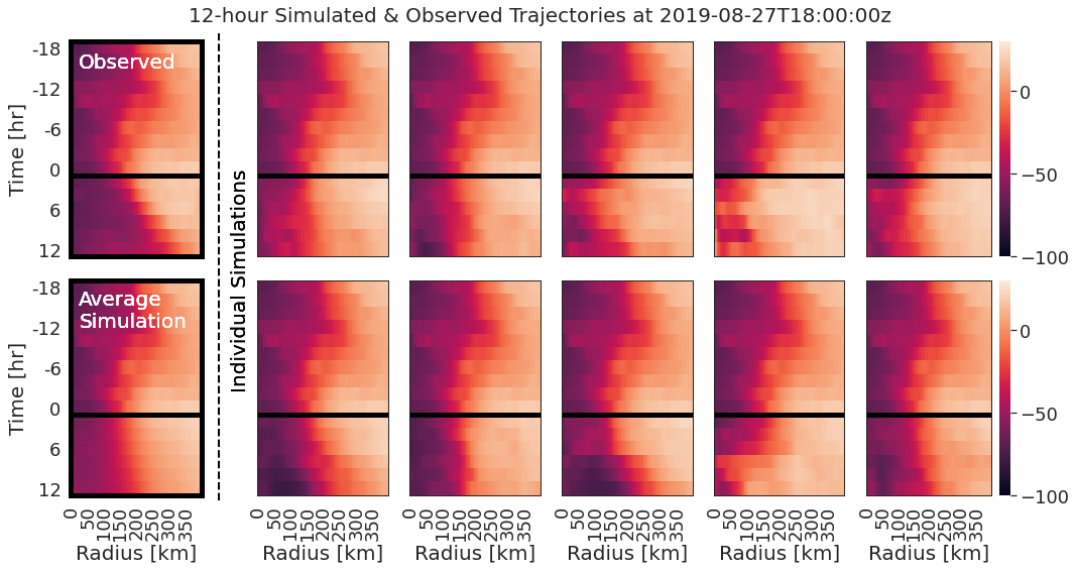


FIG. 6. **Observed and Simulated Trajectories, Hurricane Dorian [2019]:** The *observed* structural trajectory is shown in the top left corner. To the right, we see 10 *individual simulations* of radial profiles (averaged over all quadrants) at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner shows the *arithmetic mean* over 64 simulated trajectories.

two rapid intensification and two rapid weakening events. Finally, Hurricane Dorian (2019) was a powerful TC with many *in situ* observations.

(i) Intensities Figure 5 shows the 6-hour forecasts based on 64 independently simulated structural trajectories per synoptic time. Because the structural forecasts are currently based entirely on persistence—no environmental fields, such as 200-850-hPa vertical wind shear, have been included—we expect the guidance to be most useful in the short term (6- to 12-hour time frame). The steadier development of Hurricanes Jose and Dorian are well-modeled, but the swift intensity changes exhibited by Hurricane Nicole as well as the rapid intensification period of Hurricane Dorian both prove challenging to capture.

Extending lead time to 12 hours increases the variation among individual simulations, but the average simulated intensity continues to roughly track the observed intensities. The rapid intensity change events exhibited by Hurricane Nicole are challenging to forecast with only IR persistence. However, the model follows Hurricane Jose's evolution relatively well, indicating that the model has value as-is at 12-hour lead times.

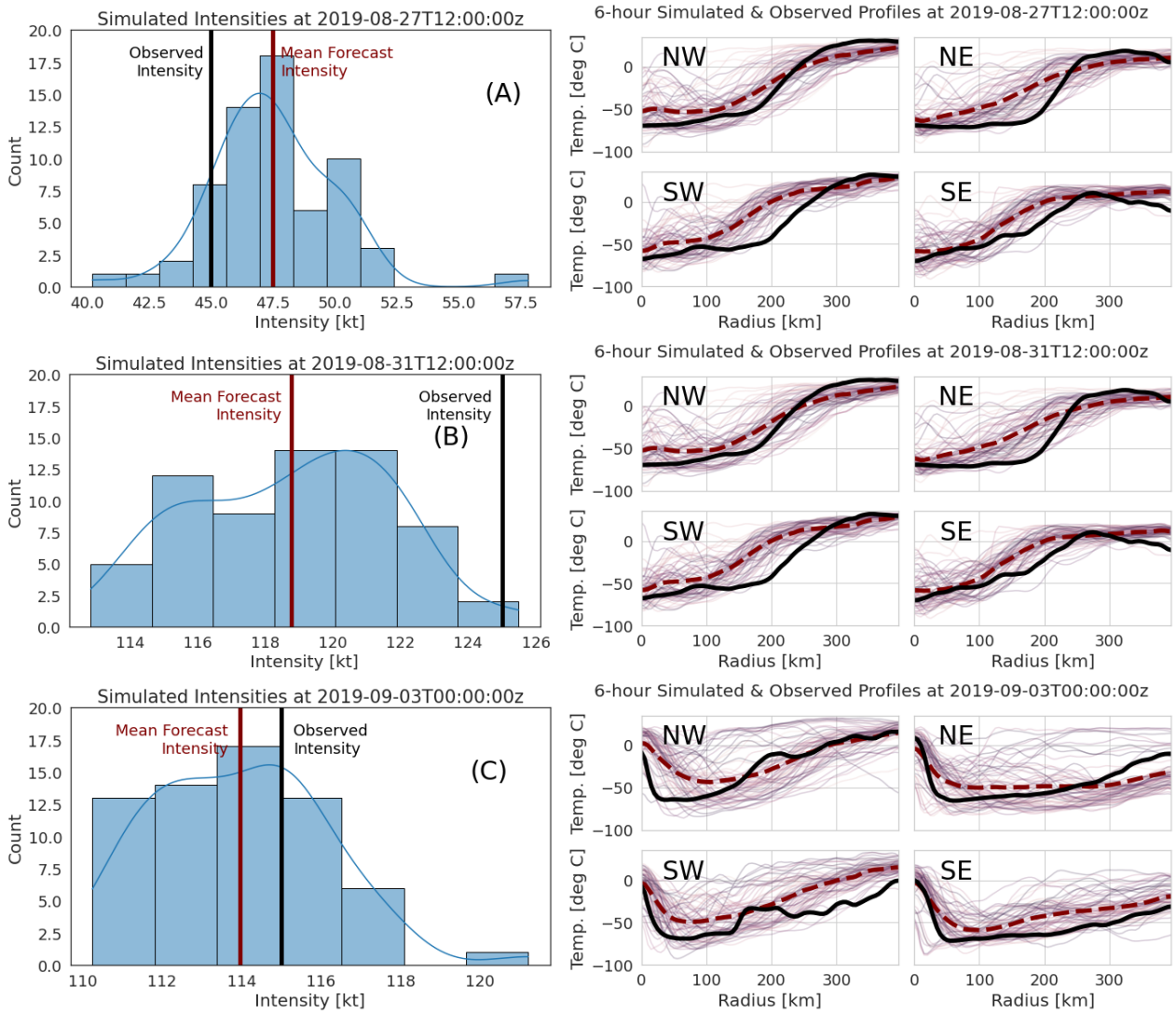


FIG. 7. **Hurricane Dorian [2019] 6-h Guidance:** (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. The distribution of the 64 intensity forecasts in the ensemble is approximated by a histogram (bar plot) and by a kernel density estimate (blue curve). (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

(ii) Diagnostics While the end goal of intensity guidance models is ultimately prediction of TC intensity, our structural forecasting pipeline adds valuable diagnostic insight into structural factors contributing to its predictions.

Figure 6 demonstrates the three-step (12-hour lead time) structural forecast for Hurricane Dorian valid for 18 UTC 27 August during a period in which it maintained 45-kt intensity. The final

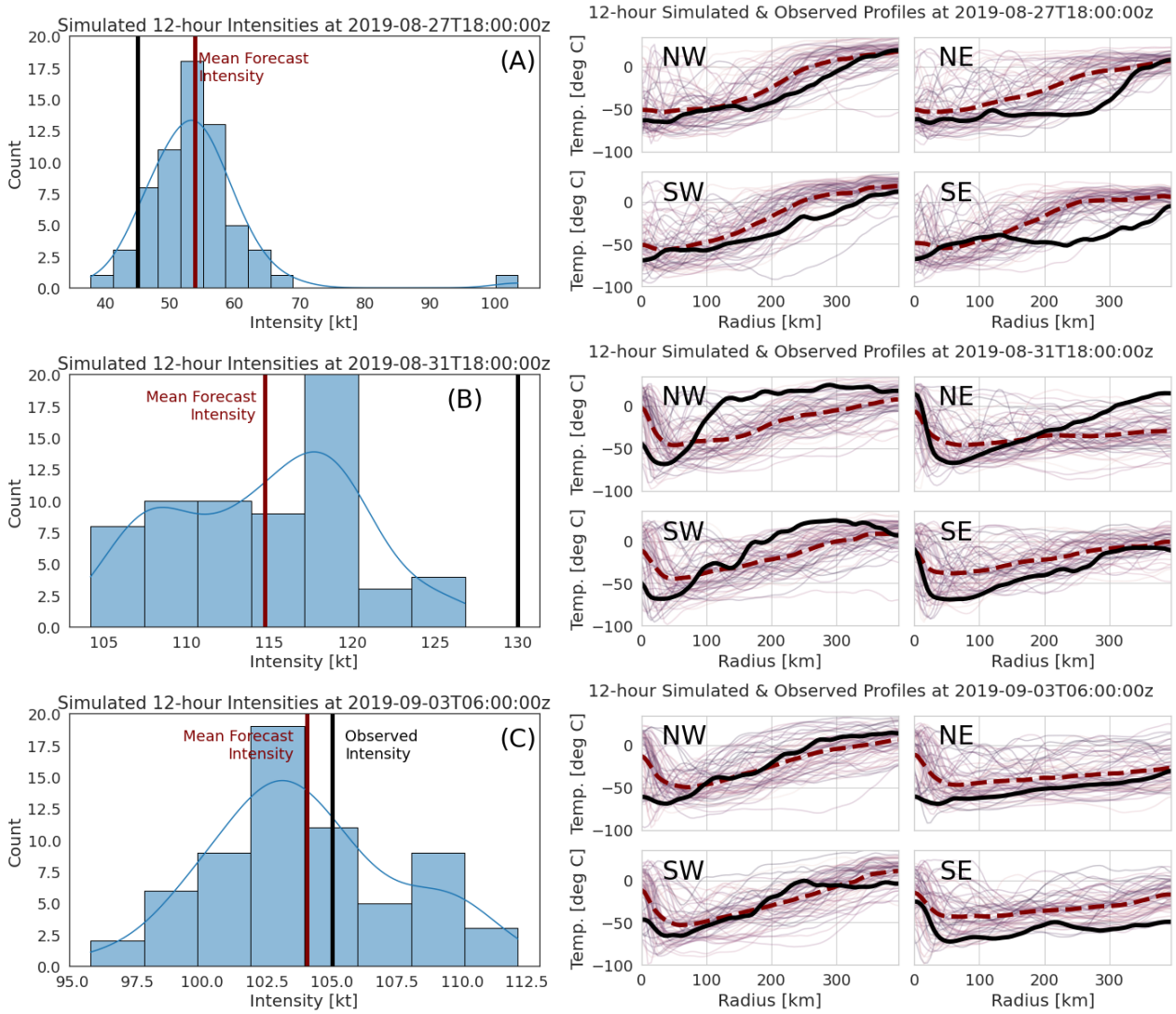


FIG. 8. **Hurricane Dorian [2019] 12-h Guidance:** Model bias during intensification (center row) is more exaggerated at 12-hour lead times. (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. The distribution of the 64 intensity forecasts in the ensemble is approximated by a histogram (bar plot) and by a kernel density estimate (blue curve). (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

6 rows of each Hovmöller diagram are simulated from the structural forecast model; in this figure, we average the four quadrants for ease of visualization (see Appendix A, Figure B10, in the supplementary materials for Hurricane Dorian structural forecasts broken down by quadrant).

Cloud-top temperature magnitude tends to be underestimated, but the expansion of cloud coverage during this 12-h period is captured across most simulations.

Figure 7 demonstrates the 6-hour forecasting guidance available at individual synoptic times. The average simulated profiles in each quadrant tend to track observed profiles reasonably well, although they tend to predict too flat a curve and too symmetric an eye. Figure 8 shows the same information but for the 12-hour lead time. Here, model biases tend to be amplified by longer lead times. We note that the emergence of an eye *is* captured in trajectory (B), even 12 hours out.

Similar figures for Hurricanes Jose and Nicole are provided in the supplemental material. In general, the structural forecast follows the observed profile, even at 12-hour lead times. We did not perform any data augmentation during training (e.g,. rotation) in order to preserve dominant geographic patterns (e.g., the prevalence of TC convection sheared eastward and northeastward in the North Atlantic), but it is possible that augmentation by rotating TCs would improve simulation fidelity, as it has been shown to improve accuracy in other TC intensity forecasting applications such as Griffin et al. (2022).

b. Model Verification

The same models are used to produce 16 simulated trajectories with associated intensity guidance for each synoptic time from 2013-2020 at the 6- and 12-hour lead times. (We use 16 rather than 64 simulations when validating over the entire 8-year period for computational reasons.) Intensity predictions provided via averaging the 16 simulations are validated against HURDAT2 best-track intensities, and past TC intensity values provided as input to the model come from operational estimates (CARQ) to emulate real-time performance.

Table 1 (left) reports the performance of the simulated trajectories versus forecast lead time in terms of root mean variance (RMV), mean absolute deviation (MAD), and bias averaged over all quadrants and radii. Let $\overline{BT}_{qi}(r)$ denote the i^{th} simulated profile in quadrant q and $\tau_q(r)$ denote

the true profile in quadrant q. Then,

$$\text{RMV} = \left(\frac{1}{400 \text{ km}} \int_0^{400} \frac{1}{4n} \sum_{q=1}^{4} \sum_{i=1}^{n} (\overline{BT}_{qi}(r) - \tau_q(r))^2 dr \right)^{\frac{1}{2}}, \tag{4}$$

$$\text{MAD} = \frac{1}{400 \text{ km}} \int_0^{400} \frac{1}{4n} \sum_{q=1}^{4} \sum_{i=1}^{n} \left| \overline{BT}_{qi}(r) - \tau_q(r) \right| dr, \tag{5}$$

$$\text{Bias} = \frac{1}{400 \text{ km}} \int_0^{400} \frac{1}{4n} \sum_{q=1}^{4} \sum_{i=1}^{n} \overline{BT}_{qi}(r) - \tau_q(r) dr. \tag{6}$$

The above are defined for simulations at a single simulation time; to combine over multiple simulation times, we average MAD and bias, and average RMV in quadrature. The measures of noise (RMV and MAD) are large even for the shortest forecast lead times; increasing the simulation size beyond 16 will reduce the impact of this noise on the average forecast. Bias, meanwhile, becomes steadily more negative with time for our structural forecasts (top left). This will lead to overestimates of intensity, as low IR temperatures are generally associated with stronger TCs. Persistence IR forecasts (bottom left) offer a less biased IR forecast on average but higher overall errors in structure at all lead times.

Table 1 (right) reports verification statistics for intensity guidance using the traditional definitions for root mean squared error (RMSE), mean absolute error (MAE), and bias. As expected, the negative bias in structural forecasts manifests as a positive bias in intensity guidance.

Tables 2 and 3 assess the performance of our intensity guidance via structural forecasting at 12-hour lead times and compare it to the NHC's official forecast verification from 2013-2019 due to availability of verification data at time of writing; note that this is a subset of the times reported in Table 1, consisting of cases where both our structural forecasts and NHC official verification are available. Overall, the RMSE of the structural forecast is about 1.1 kt larger than the NHC official forecast error as computed by RMSE, and structural forecasts produce roughly twice the bias (1.1 vs -0.6 kt). The structural forecast sees unchanged MSE with increasing 200-850-hPa vertical wind shear; the bias, however, increases with increasing wind shear (Table 2). This trend is expected, as the model does not include wind shear as a predictor but instead relies on the positive correlation between shear and asymmetry in IR imagery (as captured by radial profiles computed by quadrant). The NHC official forecast error exhibits a similar, if less pronounced, trend in bias

with increasing shear. The direction of shear seems more important, with both our model and the NHC official forecast performing most poorly for NW shear (6% of cases) and best for SW shear (9% of cases). The NE and SE cases dominate the overall model performance since they comprise the remaining 85% of the data set. The disparity between different shear magnitudes and directions could be alleviated in a model which utilizes environmental predictors.

Table 3 demonstrates similar error trends for both official forecasts and our structural forecasts. Errors tend to increase with TC intensity and with rate of intensification or weakening. The structural model produces higher bias for weaker TCs and lower bias for stronger TCs. Similarly, the structural model tends to overestimate intensities during weakening and underestimate them during intensification. The model errors are comparable to NHC official forecast errors during periods of maintenance and intensification (although bias is higher); it is periods of weakening which tend to be poorly modeled by the structural forecast. We suspect that the inclusion of environmental information could improve fidelity in weakening cases; see Section 6 on Future Work Directions for a discussion of such avenues for model improvement.

c. Variable Importance in Intensity Forecasts

Our model results show that structural forecasts result in 6-hour and 12-hour intensity predictions of comparable accuracy to NHC official forecasts. For insight into how much our model relies

Trajectory Verification

Model	Lead time	2 h	4 h	6 h	8 h	10 h	12 h
Structural	RMV	10.5C	14.7C	17.4C	22.2C	25.1C	26.9C
	MAD	7.1C	10.5C	12.8C	15.6C	17.6C	19.0C
	Bias	-0.6C	-2.2C	-4.0C	-6.3C	-7.9C	-9.3C
Persistence	RMV	12.0C	16.2C	19.7C	22.7C	28.2C	31.7C
	MAD	8.2C	11.1C	13.6C	15.8C	18.5C	20.5C
	Bias	-0.4C	0.6C	2.0C	3.5C	5.5C	6.8C

Intensity Verification

Lead time	6 h	12 h
RMSE	4.9kt	9.5kt
MAE	3.5kt	7.1kt
Bias	0.6kt	2.3kt

TABLE 1. **Overall Model Verification at n = 16:** (Left) Trajectory verification of structural forecasts, compared to IR persistence forecasts where the radial profiles are fixed at their 0 h values. Simulation noise (root mean variance and mean absolute deviation) grows rapidly in the first 6 hours; bias increases in magnitude steadily. We note that persistence offers a less biased IR forecast on average, but higher overall errors in structure at all lead times. (Right) Intensity verification vs HURDAT2 best-track intensities from 2013-2020 at each lead time.

Shear Magnitude

Shear	Structural 12-h RMSE/MAE/Bias	NHC OFCL 12-h RMSE/MAE/Bias	N
0-10kt	9.3/7.3/-0.7 kt	9.3/6.4/-1.1 kt	243
10-20kt	9.7/6.9/1.0 kt	8.1/5.7/-0.5 kt	509
20+kt	8.6/6.3/2.1 kt	7.5/5.1/-0.6 kt	439
Total	9.2/6.8/1.1 kt	8.1/5.6/-0.6 kt	1,191

Shear Direction

Shear Direction	Structural 12-h RMSE/MAE/Bias	NHC OFCL 12-h RMSE/MAE/Bias	N
SW	8.6/6.1/-0.1 kt	7.5/5.4/-1.1 kt	106
SE	9.1/6.9/0.3 kt	8.5/5.9/-0.5 kt	440
NE	9.3/6.6/2.2 kt	7.9/5.4/-0.6 kt	575
NW	10.6/8.2/-2.0 kt	8.9/6.1/-0.9 kt	70
Total	9.2/6.8/1.1 kt	8.1/5.6/-0.6 kt	1,191

TABLE 2. **Intensity Guidance Verification Relative to Shear**: Model verification binned by 200-850-hPa vertical wind shear, reported as RMSE/MAE/Bias. (Left) The performance of the structural forecasting model does not change meaningfully relative to wind shear magnitude, while the NHC official forecast performs better in higher shear environments. The structural forecast has comparable performance to the NHC official forecasts in low-shear environments. (Right) The performance of the structural model does vary with shear direction. Both the NHC forecasts and the structural model produce higher errors for NW shear (6% of cases).

TC Intensity

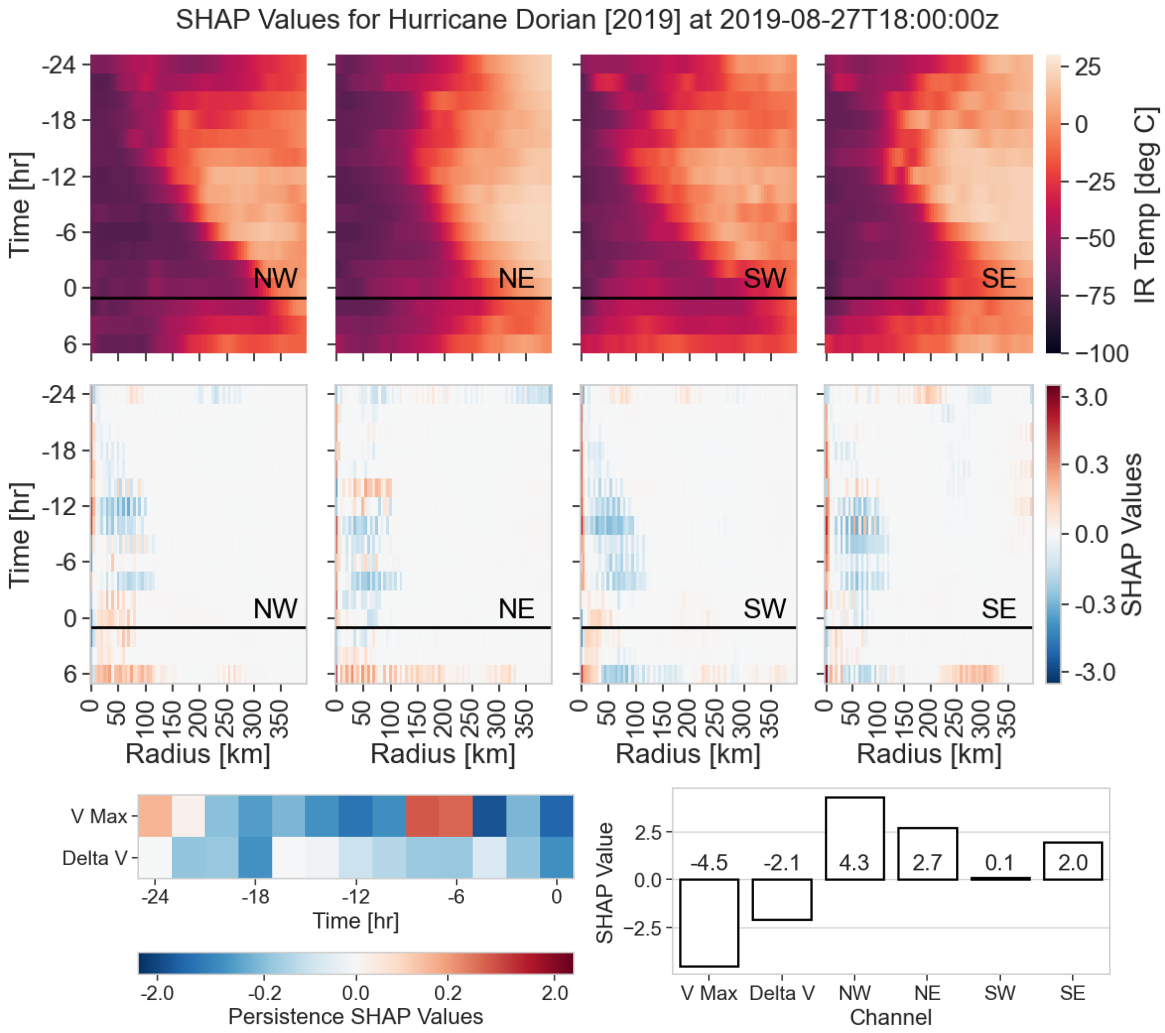
Category	Structural 12-h RMSE/MAE/Bias	NHC OFCL 12-h RMSE/MAE/Bias	N
Tropical Depression	5.5/4.4/3.0 kt	5.9/4.2/-3.1 kt	112
Tropical Storm	7.2/5.6/2.8 kt	6.4/4.4/-0.6 kt	567
Hurricane	10.3/7.7/0.6 kt	9.3/6.7/-0.6 kt	355
Major Hurricane	14.0/10.7/-5.5 kt	11.8/8.4/0.9 kt	157
Total	9.2/6.8/1.1 kt	8.1/5.6/-0.6 kt	1,208

Intensity Change

Evolution	Structural 12-h RMSE/MAE/Bias	NHC OFCL 12-h RMSE/MAE/Bias	N
Weakening	12.1/8.9/7.7 kt	7.8/5.4/2.9 kt	263
Maintenance	6.7/5.3/3.3 kt	6.1/4.3/-0.0 kt	497
Intensifying	9.8/7.3/-5.5 kt	10.2/7.2/-3.5 kt	431
Total	9.2/6.8/1.1 kt	8.1/5.6/-0.6 kt	1,208

TABLE 3. **Intensity Guidance Verification by TC Intensity**: Model verification split out by intensity and intensity change, reported as RMSE/MAE/Bias. (Left) Both the structural and NHC official forecasts struggle more with intense storms, which are rarer. The structural forecast has much stronger bias, which is expected due to the heavy influence of persistence features in the absence of environmental predictors. (Right) Similarly, both forecasts perform best during maintenance periods (6-hour change $\leq$ 5 kt in magnitude), overestimate during weakening, and underestimate during intensification. The bias is more pronounced in the structural forecast due to the absence of environmental predictors.

on IR inputs and prior intensities when making predictions, we compute a saliency map (also known as pixel attribution) for each input. There are varied definitions for saliency, including occlusion-based approaches such as SHAP explainability values (Lundberg and Lee 2017), LIME values (Ribeiro et al. 2016), and gradient-based approaches.



424 FIG. 9. **SHAP Variable Importance Map for 6-Hour Intensity Forecast:** (Top) IR channels of CNN
425 nowcasting model for Hurricane Dorian [2019] with observed IR structure above the horizontal black lines
426 and +6 h forecasted structure from deep auto-regressive generative model below the horizontal lines. (Center)
427 Pixel-wise SHAP variable importance of IR inputs on the 6-h intensity forecast. (Bottom Left) SHAP variable
428 importance of VMax (linearly interpolated operational intensity estimates) and Delta V (2-h rate of change of
429 operational intensity estimates) on the 6-h intensity forecast. (Bottom Right) Aggregated SHAP values over each
430 channel, indicating IR features contributing to intensity forecasts to a degree comparable to persistence features.
431

439 Figure 9 (center) shows a map of the SHAP importance or contribution of each pixel of the IR
440 observed and forecasted imagery on the 6-hour intensity forecast for Hurricane Dorian [2019].
441 The bottom-left panel shows the SHAP values for prior intensity and prior intensity change. The

bottom-right panel shows aggregated SHAP values for each input channel. From this result and a similar analysis with SHAP variable importance maps for Hurricane Jose [2017] and Hurricane Nicole [2016] in Appendix A and gradient-based saliency maps in Appendix B of Supplementary Materials, we conclude that (i) IR imagery contributes to the intensity forecasts to a degree comparable to persistence features, (ii) forecasted infrared imagery from our deep autoregressive generative model plays a more important role than observed past imagery in the TC intensity forecasts, (iii) the current and past presence/absence of an eye is generally the key feature of a storm, and (iv) the core temperatures outside of the eye play a signifcant role for intensity forecasting.

5. Discussion and Conclusion

This paper demonstrates a novel interpretable approach to short-term TC intensity guidance trained solely on intensity estimates up to 6 hours prior to the current time and IR observations up to 0 h. We specifically leverage spatial characteristics of TC convection as captured by radial IR profiles. By forecasting an ensemble of +6 h and +12 h trajectories of TC IR structure with radial profiles computed over four geographic quadrants, we obtain reasonable estimates of future +6 h and +12 h TC intensity while simultaneously capturing and enabling visualization of signals in convective structure relevant to those future intensities. We focus on interpretable, physically-based factors to facilitate understanding of the model's performance (e.g., upcoming intensification corresponds with decreasing cloud-top temperatures in the structural forecast). The approach outlined here has the potential for further improvement by adopting other network architectures for structural forecasts and by including environmental predictors provided in real time by SHIPS guidance. Though testing on years of cases takes time, an individual forecast for a single TC can be obtained in minutes on a single GPU, indicating the potential for the eventual use of this model as part of the available TC guidance suite in an operational setting.

6. Future Work Directions

a. Improving the Network Architecture for Structural Forecasts

The PixelSNAIL approach provides reasonable simulations of TC IR structural evolution up to 12-hour lead times. However, there exists a wealth of alternate deep autoregressive generative

models, each of which can be designed and trained in innumerable ways. Likewise, deep autoregressive models are not the only generative models available. Simulation could be carried out via vector autoregression on a low-dimensional projection of profiles (e.g., principal component analysis, Fourier bases, etc.), generative adversarial networks (GANs; Creswell et al. 2018), or transformers (e.g., temporal fusion transformers for multihorizon forecasting (Lim et al. 2021) and spatiotemporal transformers (Grigsby et al. 2021)). The PixelSNAIL architecture was chosen to demonstrate the value and feasibility of structural forecasting for intensity guidance.

b. Calibrating the Probability Distribution of Structural Forecasts

Our structural forecasts are probabilistic in nature, taking the form of *probability distributions* over future structural trajectories $\mathbf{S}_{>t}$. In the current work, we apply a standard machine learning approach of fitting a model by minimizing a loss function (in this case the negative log likelihood). A good probabilistic forecast, however, should be *conditionally calibrated*. That is, the probability of a particular event (in our case, specific radial profiles 6-12 hours into the future), given or "conditional on" a particular history of evolution and other predictors, should match the predicted probability of the same event. This is essentially saying that draws from the forecasting model should be indistinguishable from actual observations, if all relevant conditions are the same. Dey et al. (2022) recently proposed a new method for adjusting or "recalibrating" probabilistic forecasts, so that they will have his property. Indeed, one can potentially apply their procedure sequentially to each autoregressive component $p(Z_{iq}|\mathbf{Z}_{i-1},\ldots,\mathbf{Z}_1)$, for pixel $i = 1,\ldots,n$, and quadrant $q = 1,2,3,4$, so as to obtain a conditionally calibrated density over structural trajectories $\mathbf{S}_{>t}$ given present and past observations; see Discussion in Dey et al. (2022).

c. Inclusion of Environmental Variables

The PixelSNAIL model presented here is a purely autoregressive process; that is, it simulates future structural features using only past IR imagery as an input. The inclusion of environmental variables known to impact TCs such as vertical wind shear, atmospheric moisture, or sea surface temperature may improve the accuracy of the forward simulation of radial profiles, particularly of structural evolution beyond 12 hours. Such factors can be added to the PixelSNAIL architecture as additional input layers via values provided by SHIPS which are not forecasted by the model. These

inputs would then serve as "guiderails" for simulated structural evolution with potential to better capture the effects of such factors on profile asymmetry. Despite these limitations, our prototype model (which is derived solely from prior and present TC intensity estimates and Geo IR imagery alongside forecasted TC structure using a very simple network architecture) provides reasonable short-term structural and intensity forecasts comparable to NHC forecasts at 6- and 12-h lead times. The inclusion of environmental variables in the nowcasting model is likely to improve its intensity forecasts, which would then be compared to SHIPS forecasts as well as NHC official forecasts, the latter of which are crafted using SHIPS and other guidance.

Acknowledgments. Part of this research was done as an Independent Study in Spring 2021 while Pavel Khokhlov was a Master in Machine Learning student at Carnegie Mellon University. We are grateful to Microsoft for providing Azure computing resources for this work. The authors would like to thank Katerina Fragkiadaki for a discussion on deep generative networks, and Galen Vincent for many helpful comments on the research. This work is supported in part by NSF DMS-2053804, NSF PHY-2020295, and the C3.ai Digital Transformation Institute.

Data availability statement. Code to generate Geo IR radial profiles from openly available data can be found at `https://github.com/ihmcneely/ORB2sample`, which draws from the MERGIR database openly available from NASA at `https://disc.gsfc.nasa.gov/datasets/GPM_MERGIR_1/summary`. The HURDAT2 best track database is openly available from the NHC at `https://www.nhc.noaa.gov/data`, while the ATCF operational best tracks (B-deck) are openly available from the National Center for Atmospheric Research at `http://hurricanes.ral.ucar.edu/repository/`. Official forecast verification files are openly available from the NHC at `https://www.nhc.noaa.gov/verification/`. Finally, SHIPS developmental data is openly available from the Cooperative Institute for Research in the Atmosphere at `https://rammb.cira.colostate.edu/research/tropical_cyclones/ships/developmental_data.asp`.

References

Chen, X., N. Mishra, M. Rohaninejad, and P. Abbeel, 2018: Pixelsnail: An improved autoregressive generative model. *International Conference on Machine Learning*, PMLR, 864–872.

Combinido, J. S., J. R. Mendoza, and J. Aborot, 2018: A convolutional neural network approach for estimating tropical cyclone intensity using satellite-based infrared images. *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 1474–1480.

Creswell, A., T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, 2018: Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, **35 (1)**, 53–65.

DeMaria, M., 2018: SHIPS developmental database file format and predictor descriptions (2018). Tech. rep., Colorado State University. URL http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/docs/ships_predictor_file_2018.doc, Developmental Data.

DeMaria, M., and J. Kaplan, 1999: An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and Eastern North Pacific basins. *Weather and Forecasting*, **14 (3)**, 326–337.

DeMaria, R., 2015: Automated tropical cyclone eye detection using discriminant analysis. Ph.D. thesis, Colorado State University.

Dey, B., D. Zhao, J. A. Newman, B. H. Andrews, R. Izbicki, and A. B. Lee, 2022: Calibrated predictive distributions via diagnostics for conditional coverage. *arXiv preprint arXiv:2205.14568*.

Dvorak, V. F., 1975: Tropical cyclone intensity analysis and forecasting from satellite imagery. *Monthly Weather Review*, **103 (5)**, 420–430, https://doi.org/10.1175/1520-0493(1975)103⟨0420:TCIAAF⟩2.0.CO;2.

Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, **101 (12)**, E2149–E2170.

Gray, W. M., 1979: Hurricanes: Their formation, structure and likely role in the tropical circulation. *Meteorology over the tropical oceans*, **155**, 218.

Griffin, S. M., A. Wimmers, and C. S. Velden, 2022: Predicting rapid intensification in north atlantic and eastern north pacific tropical cyclones using a convolutional neural network. *Weather and Forecasting*.

Grigsby, J., Z. Wang, and Y. Qi, 2021: Long-range transformers for dynamic spatiotemporal forecasting. *arXiv preprint arXiv:2109.12218*.

Hovmöller, E., 1949: The trough-and-ridge diagram. *Tellus*, **1 (2)**, 62–66.

Hu, L., E. A. Ritchie, and J. S. Tyo, 2020: Short-term tropical cyclone intensity forecasting from satellite imagery based on the deviation angle variance technique. *Weather and Forecasting*, **35 (1)**, 285–298.

Janowiak, J., B. Joyce, and P. Xie, 2020: NCEP/CPC L3 half hourly 4km global (60S - 60N) merged IR v1. NASA Goddard Earth Sciences Data and Information Services Center, URL https://disc.gsfc.nasa.gov/datacollection/GPM_MERGIR_1.html, https://doi.org/10.5067/P4HZB9N27EKU.

Knaff, J. A., and R. T. DeMaria, 2017: Forecasting tropical cyclone eye formation and dissipation in infrared imagery. *Weather and Forecasting*, **32 (6)**, 2103–2116.

Knapp, K. R., and S. L. Wilkins, 2018: Gridded satellite (GridSat) GOES and CONUS data. *Earth System Science Data*, **10 (3)**, 1417–1425, https://doi.org/10.5194/essd-10-1417-2018.

Lee, J., J. Im, D.-H. Cha, H. Park, and S. Sim, 2019: Tropical cyclone intensity estimation using multi-dimensional convolutional neural networks from geostationary satellite data. *Remote Sensing*, **12 (1)**, 108.

Lim, B., S. Ö. Arık, N. Loeff, and T. Pfister, 2021: Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, **37 (4)**, 1748–1764.

Liu, Z., 2021: private communication.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Advances in neural information processing systems*, **30**.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175–2199.

McNeely, T., A. B. Lee, D. Hammerling, and K. Wood, 2019: Quantifying the spatial structure of tropical cyclone imagery.

McNeely, T., A. B. Lee, K. M. Wood, and D. Hammerling, 2020: Unlocking GOES: A statistical framework for quantifying the evolution of convective structure in tropical cyclones. *Journal of Applied Meteorology and Climatology*, **59 (10)**, 1671–1689.

McNeely, T., G. Vincent, A. B. Lee, R. Izbicki, and K. M. Wood, 2022: Detecting distributional differences in labeled sequence data with application to tropical cyclone satellite imagery. *arXiv preprint arXiv:2202.02253*.

Olander, T., A. Wimmers, C. Velden, and J. P. Kossin, 2021: Investigation of machine learning using satellite-based advanced dvorak technique analysis parameters to estimate tropical cyclone intensity. *Weather and Forecasting*, **36 (6)**, 2161–2186.

Olander, T. L., and C. S. Velden, 2007: The advanced dvorak technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Weather and Forecasting*, **22 (2)**, 287–298.

Pradhan, R., R. S. Aygun, M. Maskey, R. Ramachandran, and D. J. Cecil, 2017: Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing*, **27 (2)**, 692–702.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Salimans, T., A. Karpathy, X. Chen, and D. P. Kingma, 2017: Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*.

Sanabia, E. R., B. S. Barrett, and C. M. Fine, 2014: Relationships between tropical cyclone intensity and eyewall structure as determined by radial profiles of inner-core infrared brightness temperature. *Monthly Weather Review*, **142 (12)**, 4581–4599, https://doi.org/10.1175/MWR-D-13-00336.1.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bulletin of the American Meteorological Society*, **98 (4)**, 681–698.

Tian, W., W. Huang, L. Yi, L. Wu, and C. Wang, 2020: A cnn-based hybrid model for tropical cyclone intensity estimation in meteorological industry. *IEEE Access*, **8**, 59 158–59 168.

Van den Oord, A., N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and Coauthors, 2016: Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, **29**.

Van Oord, A., N. Kalchbrenner, and K. Kavukcuoglu, 2016: Pixel recurrent neural networks. *International conference on machine learning*, PMLR, 1747–1756.

Wang, C., Q. Xu, X. Li, and Y. Cheng, 2020: Cnn-based tropical cyclone track forecasting from satellite infrared images. *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 5811–5814.

Zhang, C.-J., X.-J. Wang, L.-M. Ma, and X.-Q. Lu, 2021: Tropical cyclone intensity classification and estimation using infrared satellite images with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **14**, 2070–2086.

Supplemental Material for "Structural Forecasting for Short-term Tropical Cyclone Intensity Guidance"

Trey McNeely,[a,d] Pavel Khokhlov, Niccolò Dalmasso,[b] Kimberly M. Wood,[c] and Ann B. Lee[a]

[a] *Carnegie Mellon University Department of Statistics and Data Science*

[b] *J.P. Morgan AI Research*

[c] *Mississippi State University Department of Geosciences*

[d] *Microsoft AI Development Acceleration Program*

Corresponding author: Ann B. Lee, annlee@stat.cmu.edu

APPENDIX A

Additional SHAP Variable Importance Maps for TC Intensity Forecasts

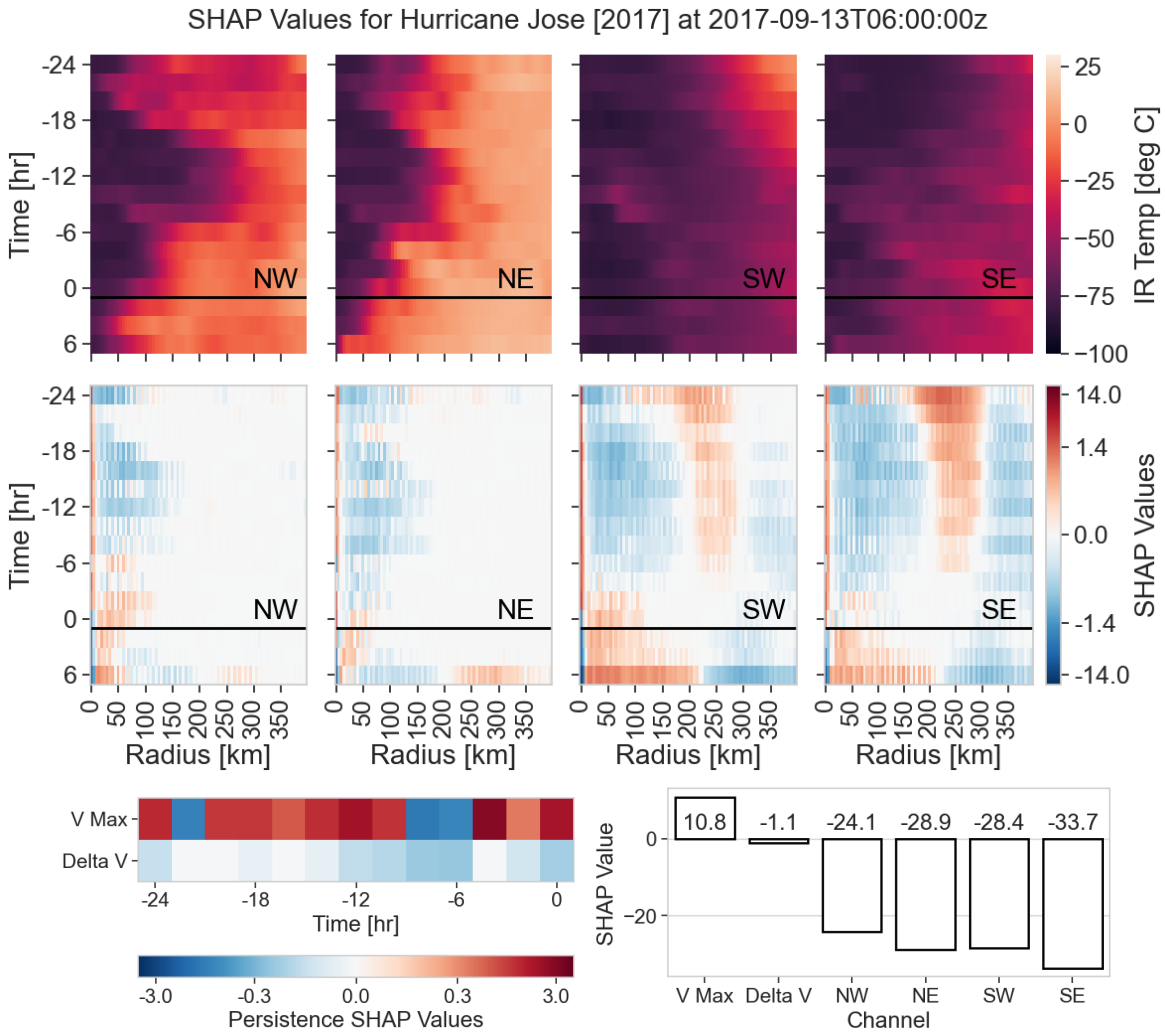


FIG. A1. **SHAP Variable Importance Map for 6-Hour Intensity Forecast:** (Top) IR channels of CNN "nowcasting" model for Hurricane Jose [2017] with observed IR structure above the horizontal black lines, and +6 hour forecasted structure from deep auto-regressive generative model below the horizontal lines. (Center) Pixel-wise SHAP variable importance of IR inputs on 6-hour intensity forecast. (Bottom Left) SHAP variable importance of VMax (linearly interpolated operational intensity estimates) and Delta V (2-h rate of change of operational intensity estimates) on the 6-h intensity forecast. (Bottom Right) Aggregated SHAP values over each channel showing IR features contributing significantly to the intensity forecasts.

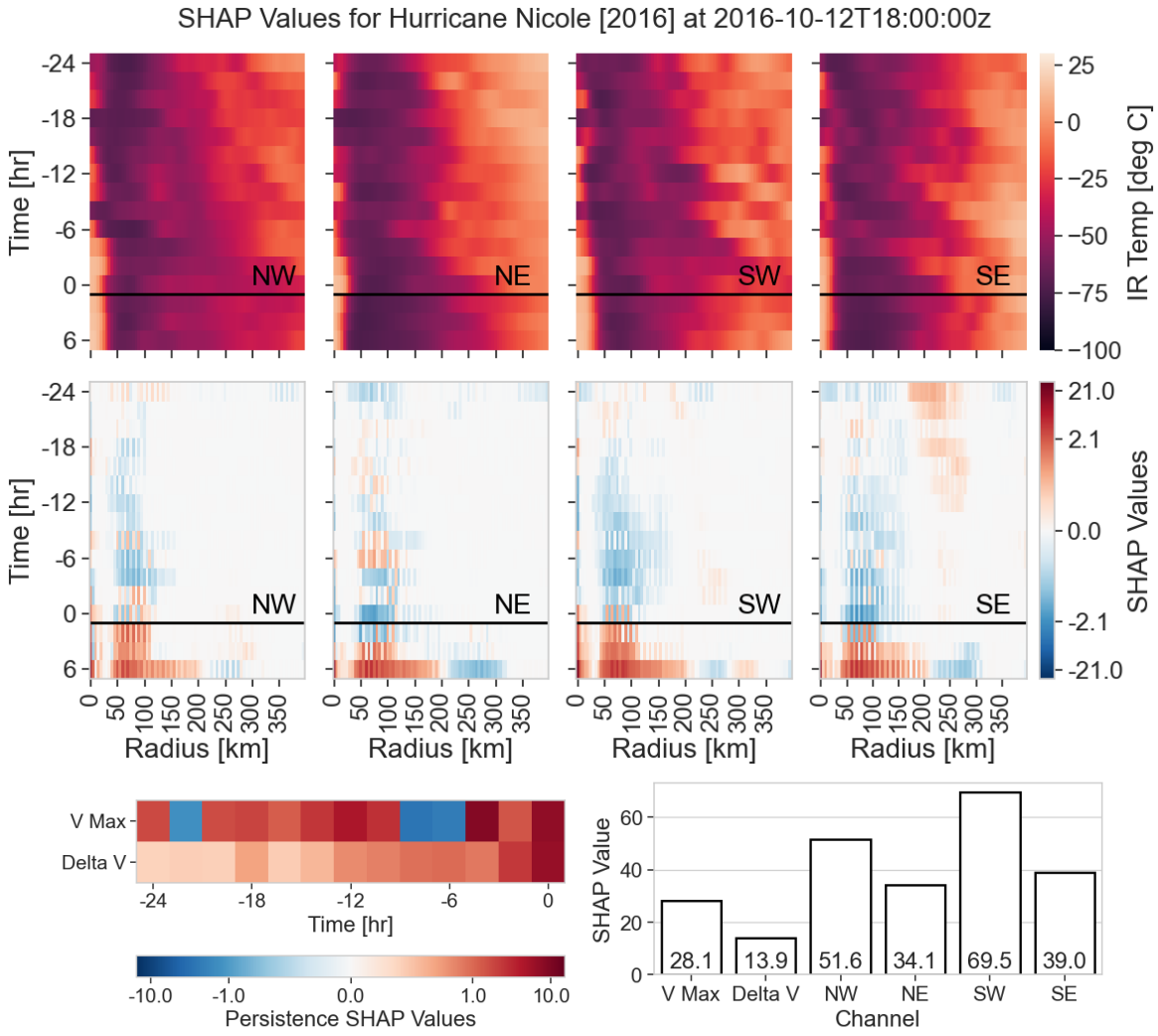


FIG. A2. **IR Channels and SHAP Variable Importance Map for 6-Hour Intensity Forecast:** As figure A1, but for Hurricane Nicole [2016].

APPENDIX B

Input Saliency for Forecasting Model

Our model results in Section 4 showed that structural forecasts result in 12-hour intensity predictions of comparable accuracy to NHC official forecasts. Here we apply a simple gradient-based approach to provide some insight as to how much the model relies on different IR inputs when making predictions. The gradient describes how much a feature contributes to the model response Y. More specifically, we define the saliency $S_i(\mathbf{x})$ of the i^{th} pixel or feature by

$$S_i(\mathbf{x}) = \left| \frac{\partial Y}{\partial x_i}\Big|_{\mathbf{x}} \right|, \tag{B1}$$

where $\mathbf{x}$ denotes the total input.

In order to visualize the overall impact of each input channel (four IR quadrants, the radius channel, the time channel, the observed prior intensity, and the observed prior intensity change) on the forecasted future intensity, we "aggregate" the saliency, summing over all pixels in each channel. Figure B1 shows the saliency aggregated by channel over time for each of our three example TCs. Note that because prior intensity/intensity change are not included in the convolutional layers, they are linear, and thus have a fixed saliency; because the model is nonlinear in the other channels, the saliency varies over time with the IR inputs. Of particular note is that *the aggregated saliency of the IR input channels is comparable to the persistence features*, indicating that the model does not simply rely on persistence to make its predictions but instead makes use of the structural forecasts. Figure B2 shows the same values for IR channels with the mean and trend removed, demonstrating that the model tends to rely more heavily on convective structure in the southern quadrants, and particularly the southeast quadrant.

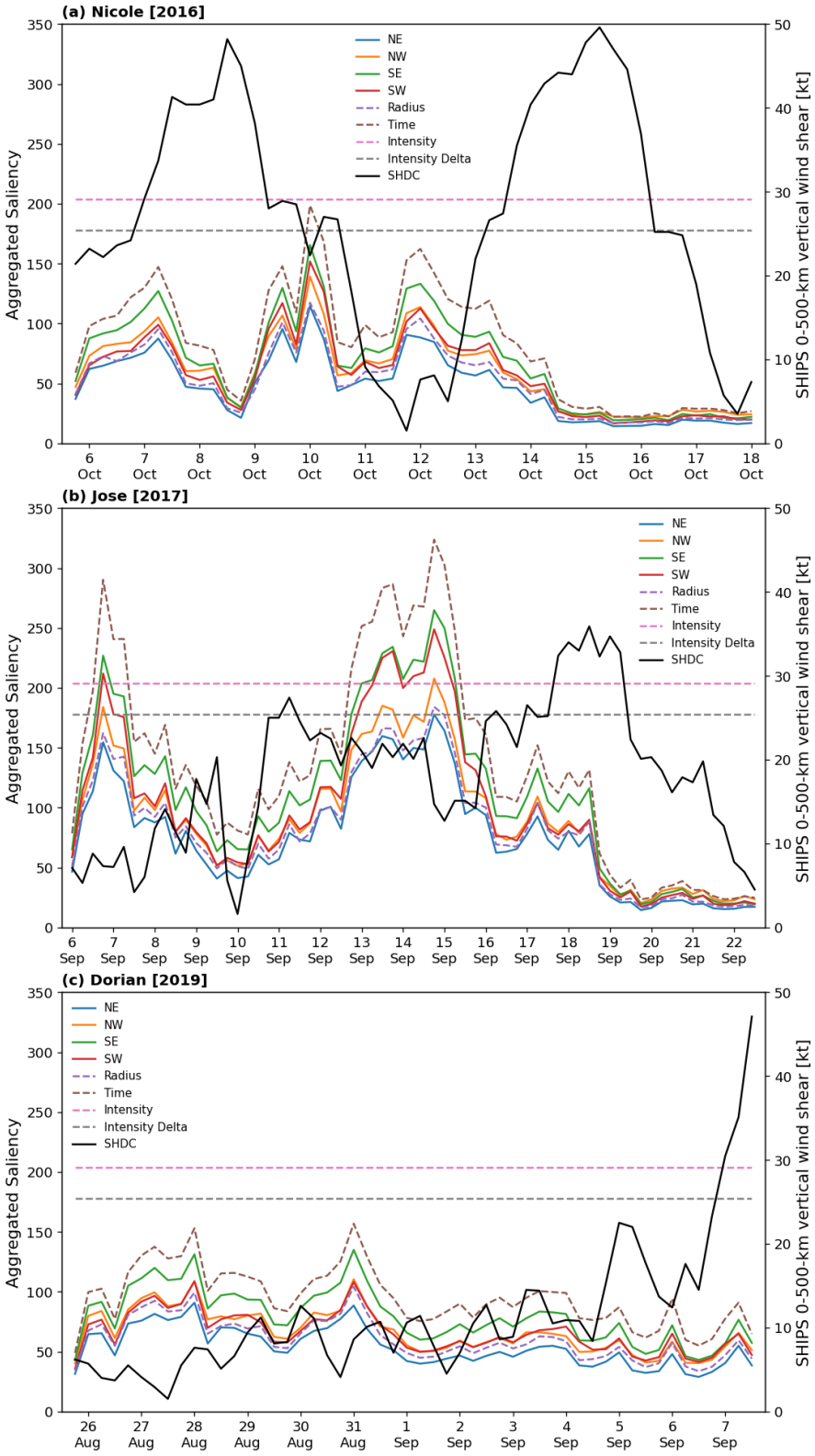


FIG. B1. Saliency per input layer in the model. Saliency values over time indicate that the model for TC intensity forecasts utilizes *image* inputs to a degree comparable to prior intensity values.

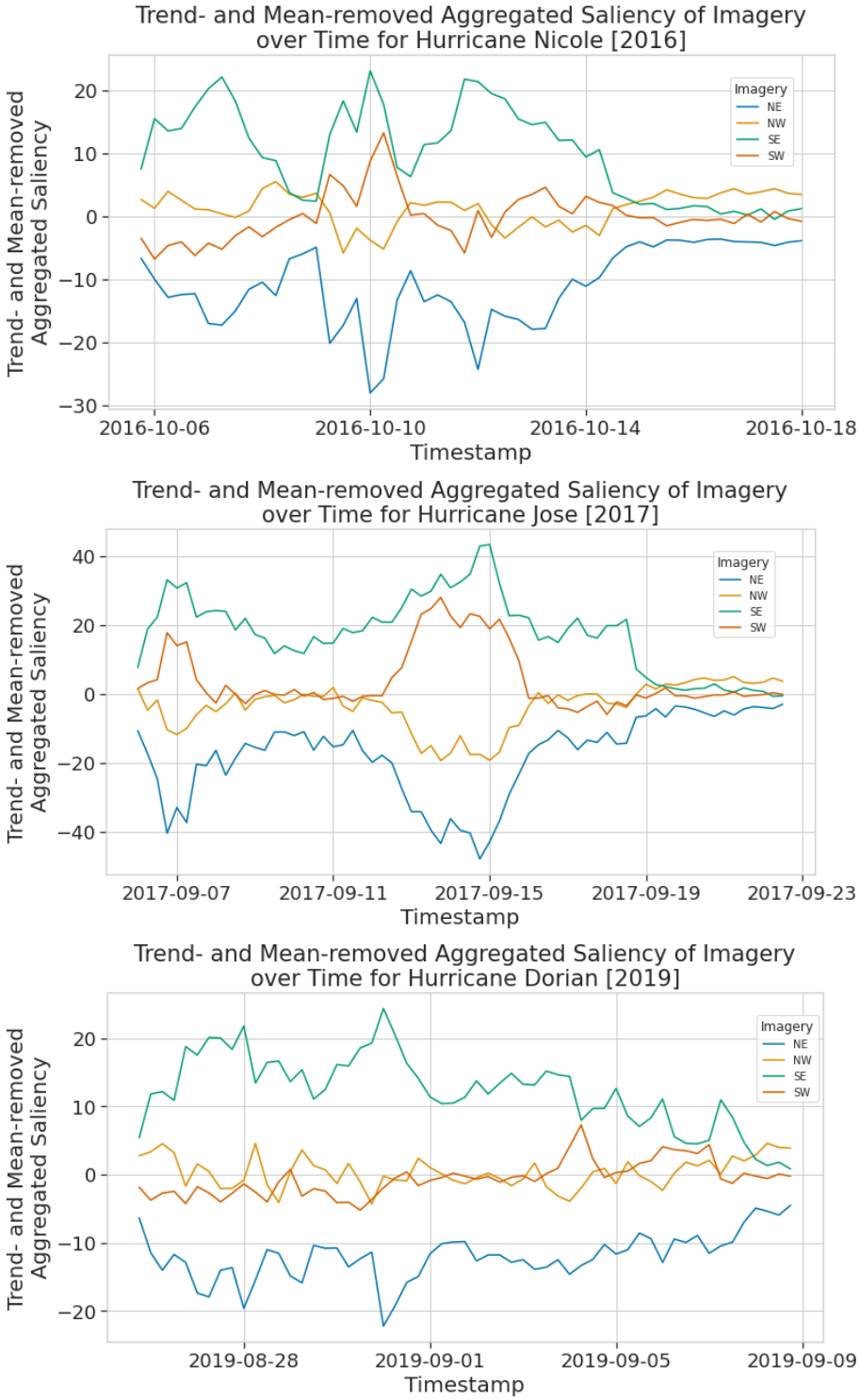


FIG. B2. Trend- and mean-removed saliency for image inputs. Comparing the saliency values over time indicates 1) the southern quadrants are more heavily utilized by the forecasting model, and 2) the degree to which the model relies on one quadrant over another is not constant (e.g. Hurricane Jose [2017] around Sept. 14, 2017).

APPENDIX C

Additional Forecast Materials for Case Studies

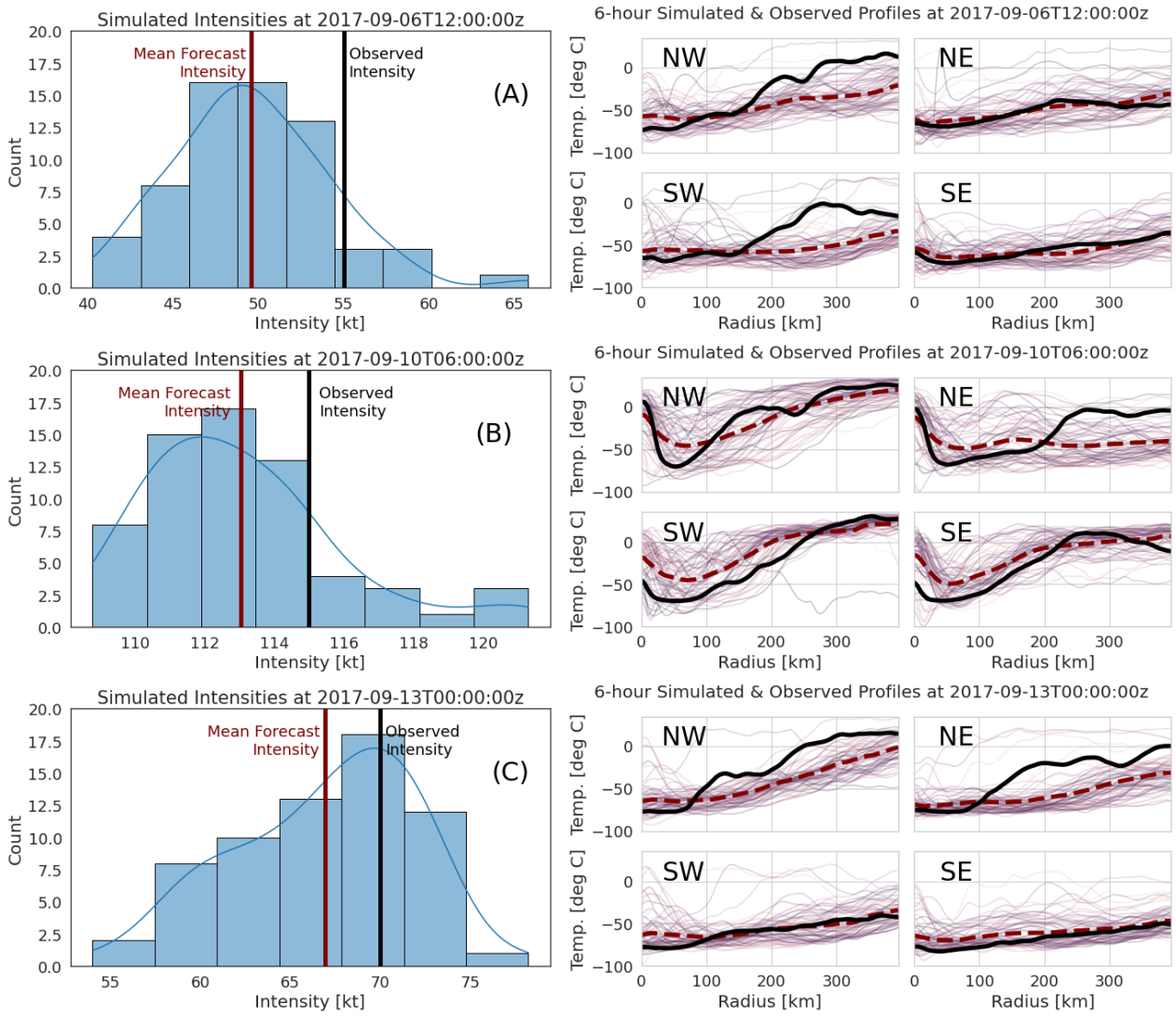


FIG. C1. **Hurricane Jose [2017] 6-hr Guidance:** Hurricane Jose was subjected to vertical wind shear out of the west/northwest due to Hurricane Irma; the structural forecasts tend to underestimate temperatures in the NW quadrant and thus overestimate TC intensity. (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

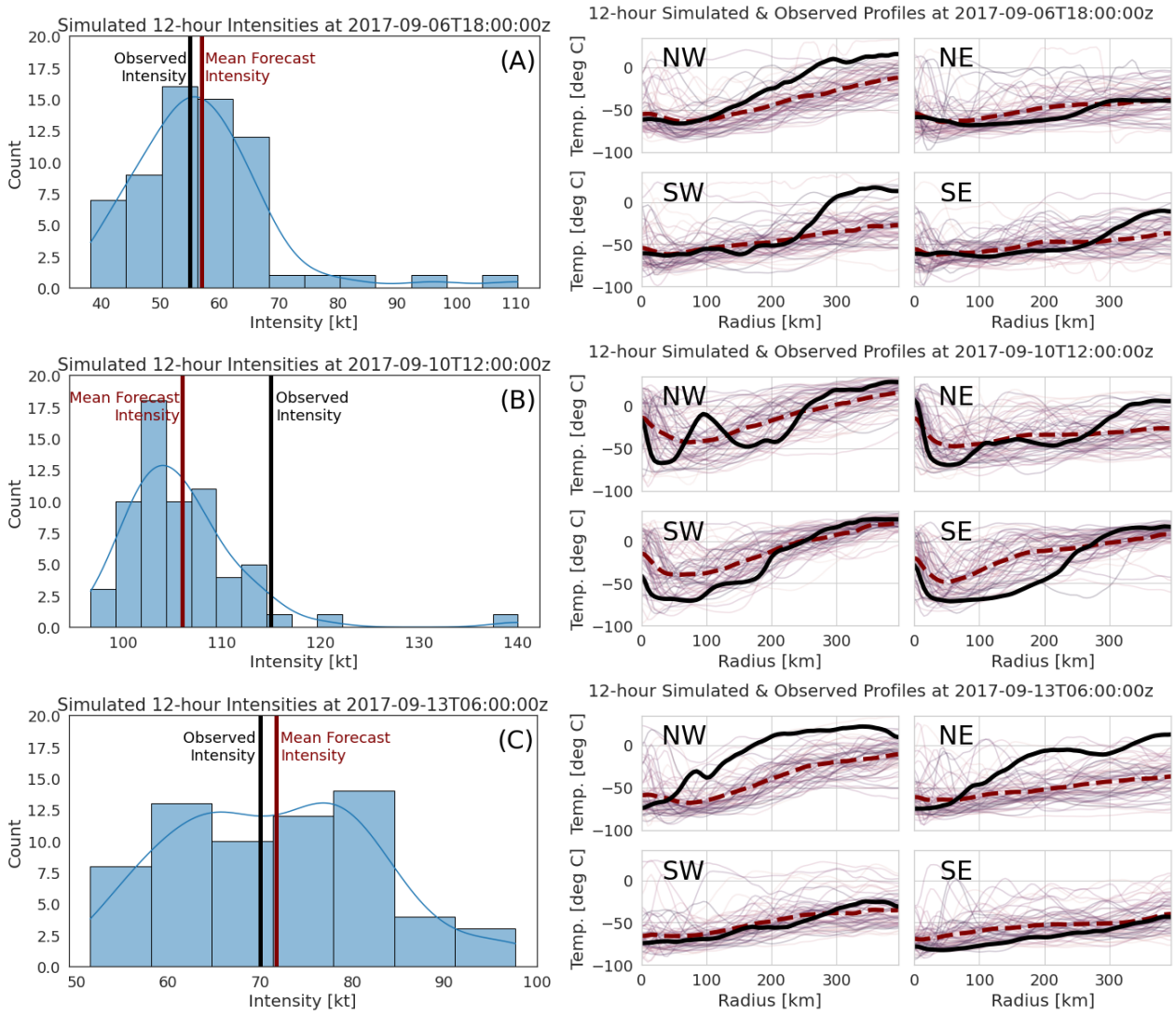


FIG. C2. **Hurricane Jose [2017] 12-hr Guidance:** At 12-hour lead times, this TC's evolution is still well modeled save for a handful of individual overestimates. (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

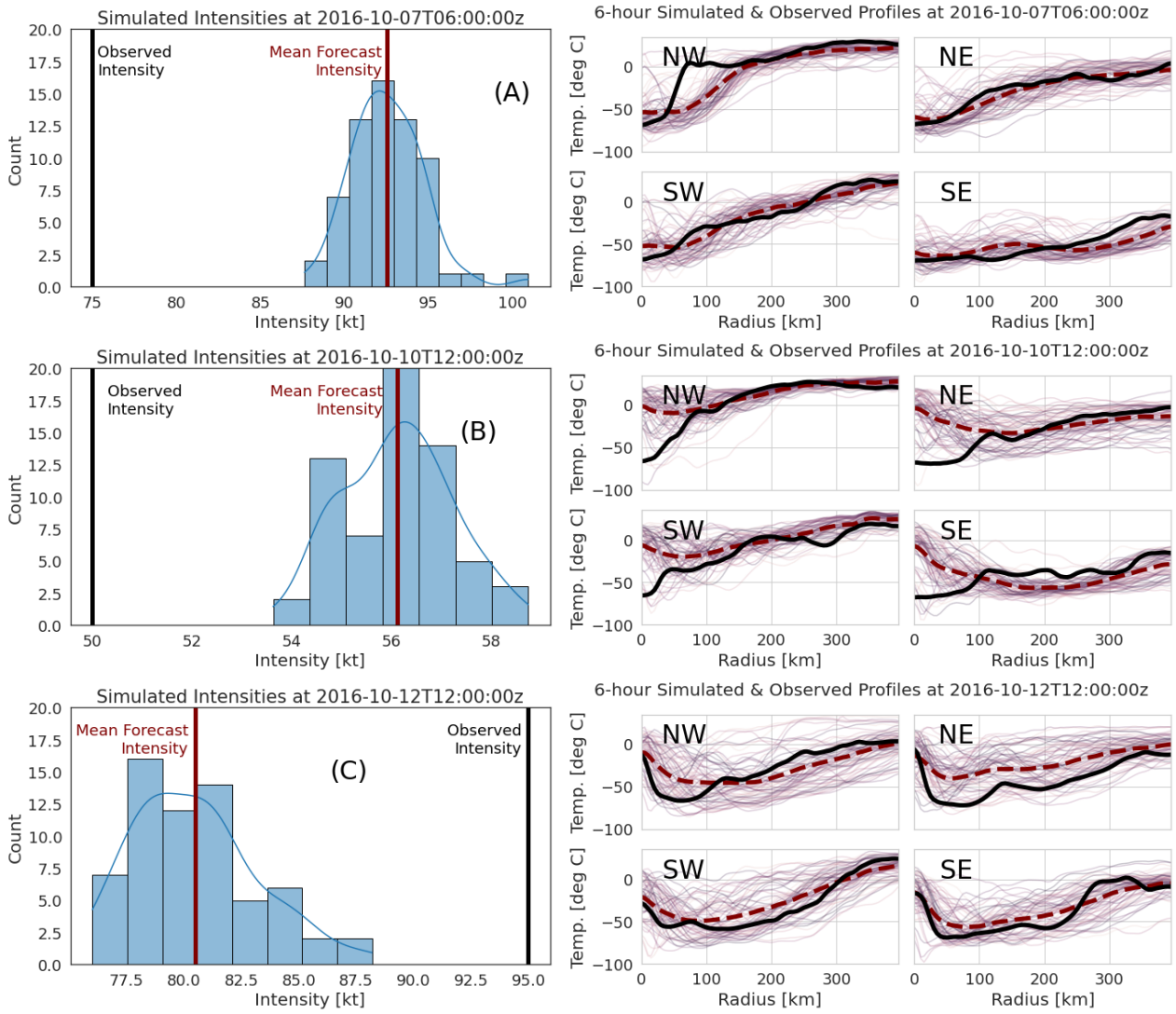


FIG. C3. **Hurricane Nicole [2016] 6-hr Guidance:** Hurricane Nicole underwent several rapid intenity change events. At 6 UTC 7 October, the structural forecast models the cloud top temperatures well, but this is insufficient to predict the extreme change from intensification to weakening even at 6-hour lead times. (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

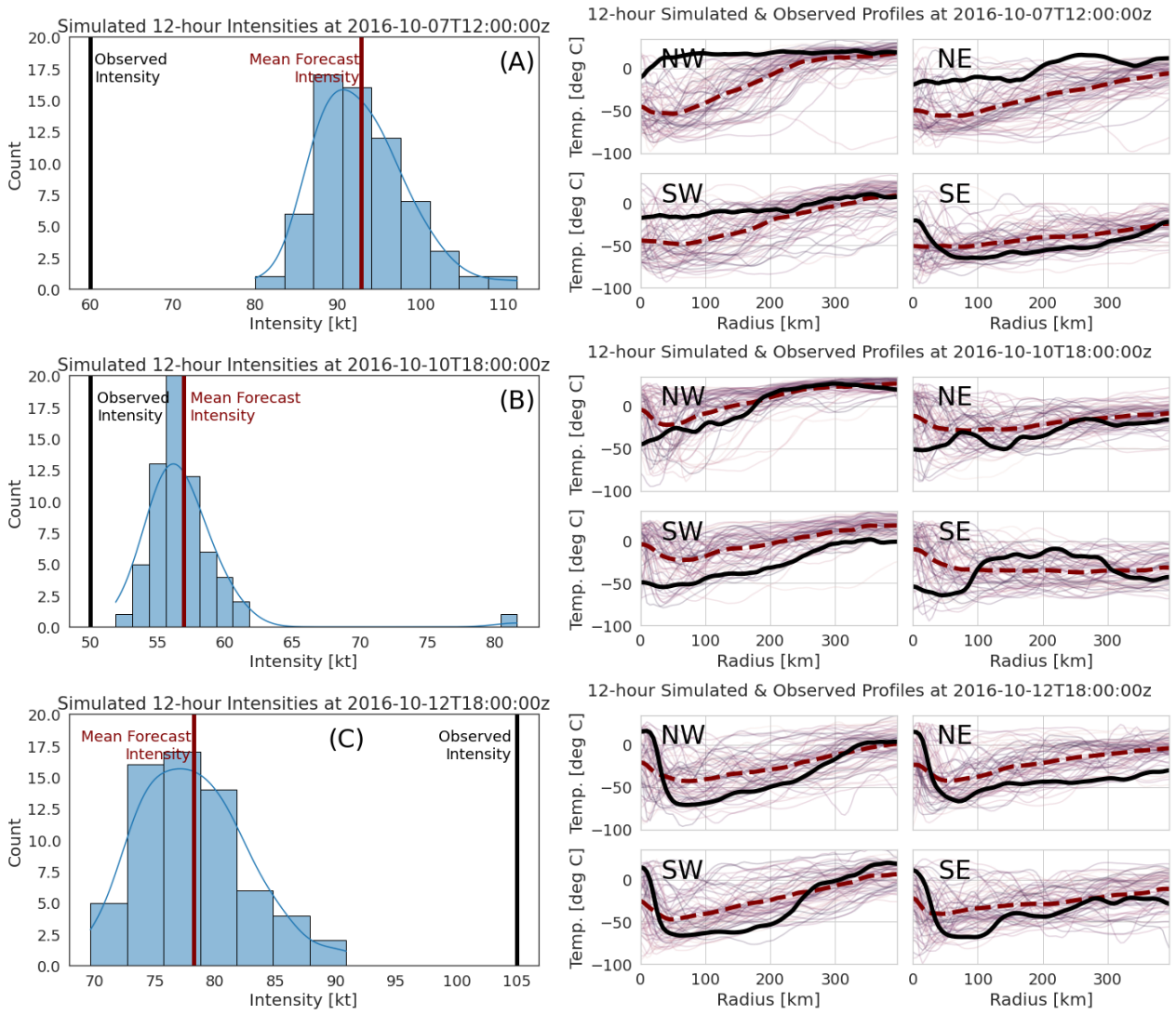


FIG. C4. **Hurricane Nicole [2016] 12-hr Guidance:** Model biases during the quick shifts from intensification to weakening and during stead intensification are exacerbated at 12-hour lead times. (Left) Distribution of forecasted intensities with observed (black) and average forecast (red) intensities marked. (Right) Simulated profiles by quadrant with observed profiles represented by solid black curves, and averaged simulated profiles represented by dashed red curves.

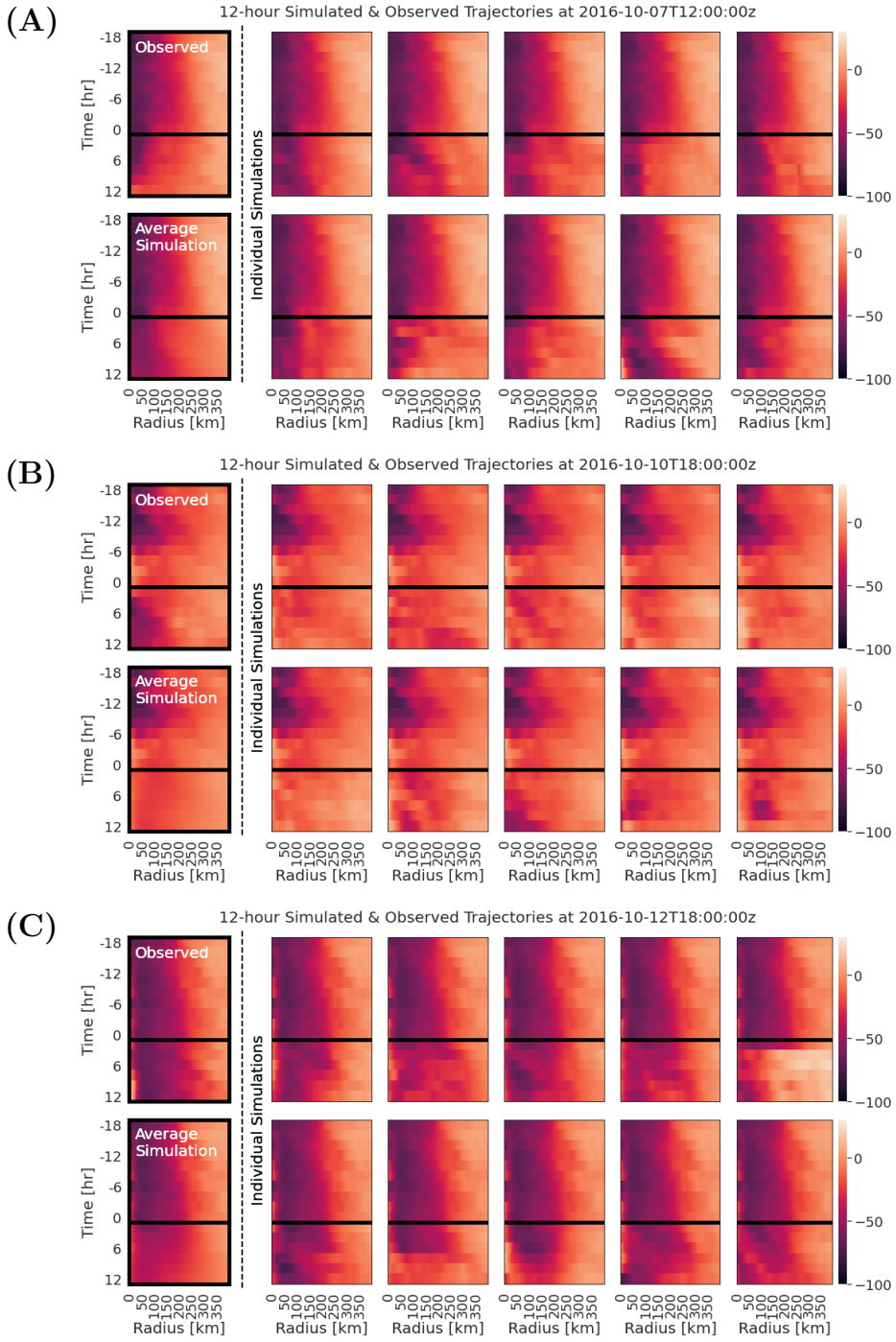


FIG. C5. **Hurricane Nicole [2016] 12-hr Structural Forecasts:** The *observed* structural trajectory is shown in the top left corner of each row. To the right, we see 10 *individual simulations* of radial profiles (averaged over all quadrants) at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner shows the *average simulation* over 64 simulated trajectories.

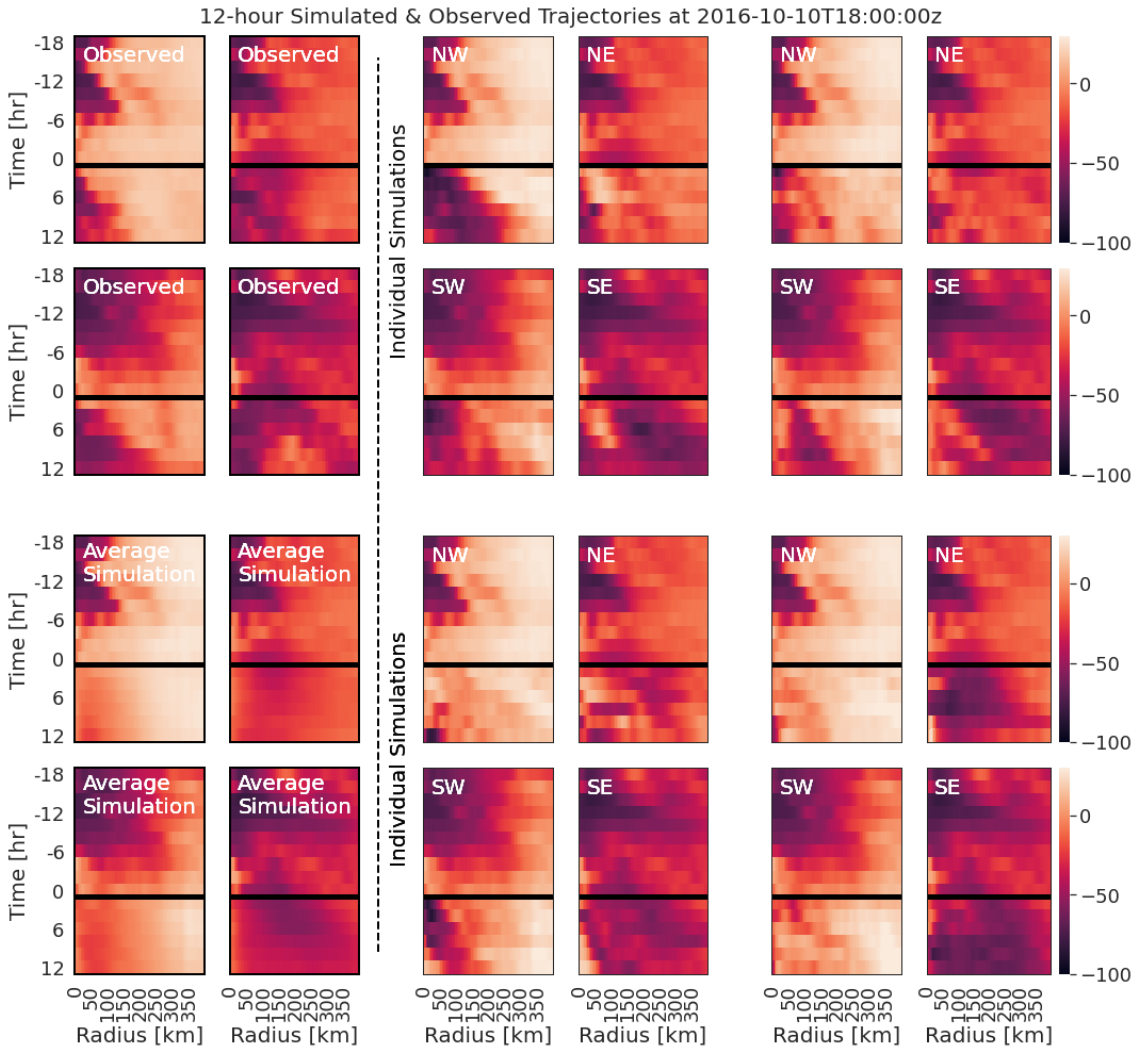
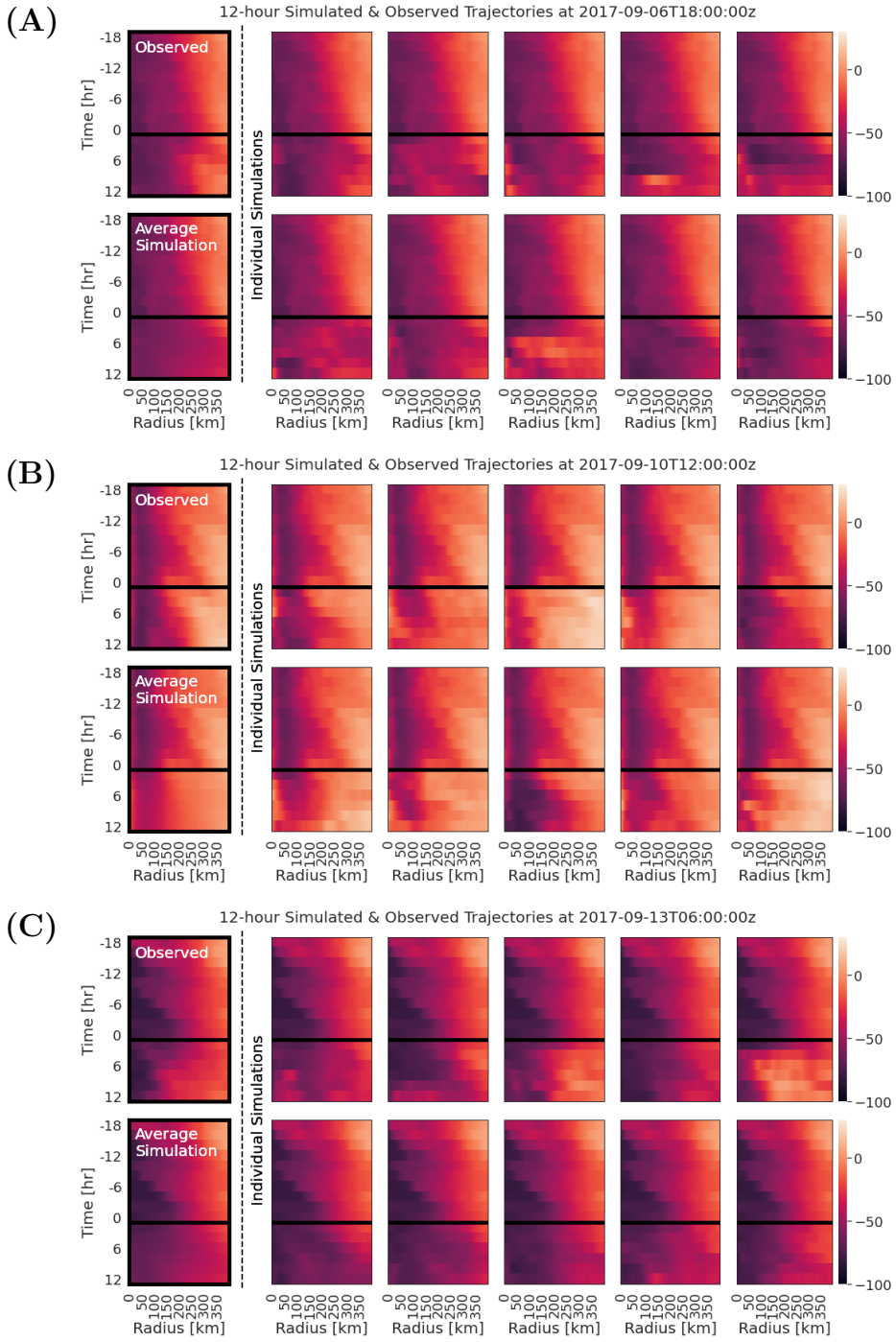


FIG. C6. **Hurricane Nicole [2016] Forecasts By Quadrant (B):** As Figure C5, but broken down by quadrant for example (B) only. The *observed* structural trajectory is shown in the top left corner. To the right, we see four *individual simulations* of radial profiles by quadrant at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner shows the *average simulation* over 64 simulated trajectories.



78 FIG. C7. **Hurricane Jose [2017] 12-hr Structural Forecasts:** The *observed* structural trajectory is shown in
79 the top left corner of each row. To the right, we see 10 *individual simulations* of radial profiles (averaged over all
80 quadrants) at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles
81 below the black line are simulated. The bottom left corner, shows the *average simulation* over 64 simulated
82 trajectories.

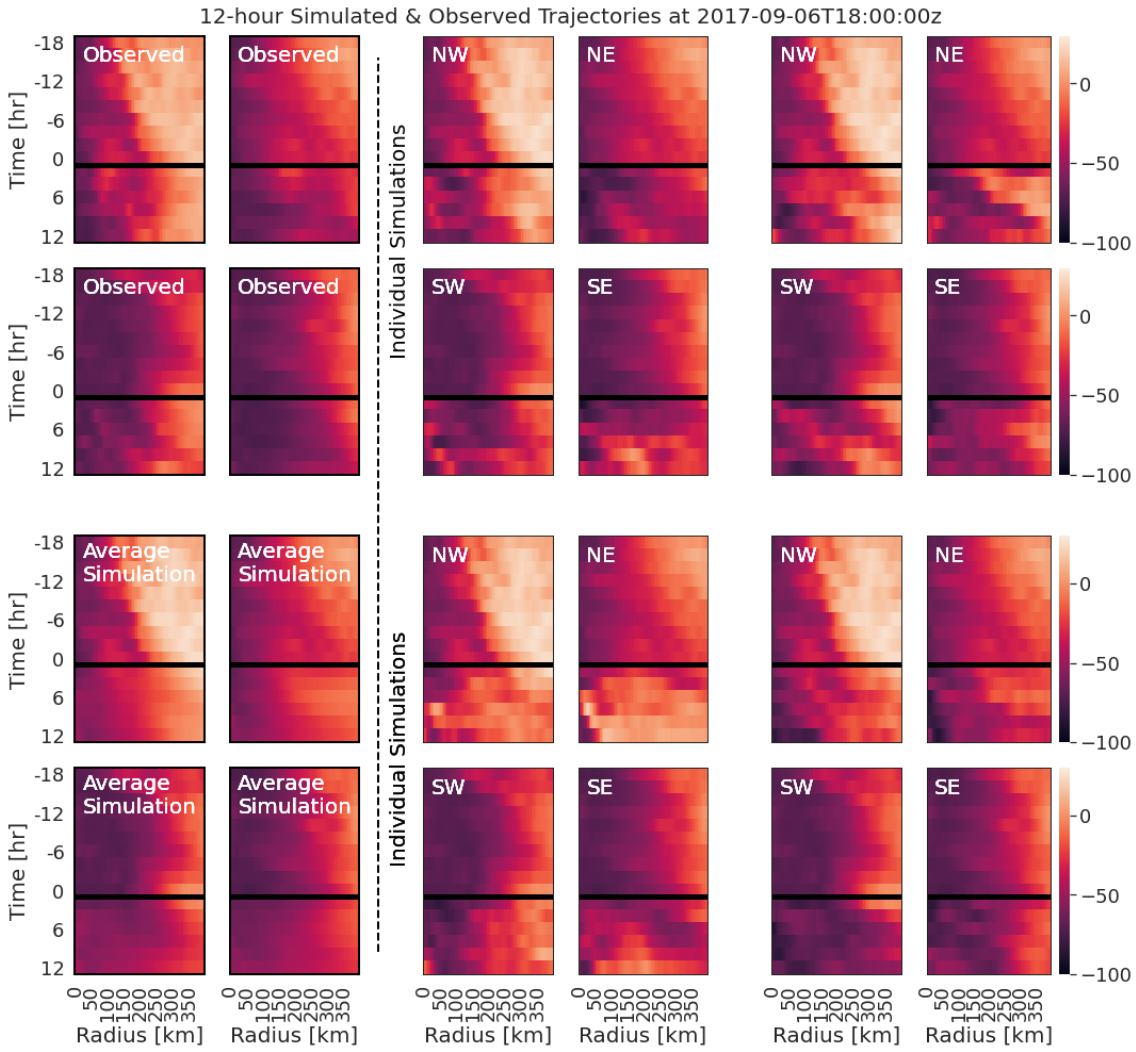


FIG. C8. **Hurricane Jose [2017] Forecasts By Quadrant (A):** As Figure C7, but broken down by quadrant for example (A) only. The *observed* structural trajectory is shown in the top left corner. To the right, we see four *individual simulations* of radial profiles by quadrant at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner shows the *average simulation* over 64 simulated trajectories.

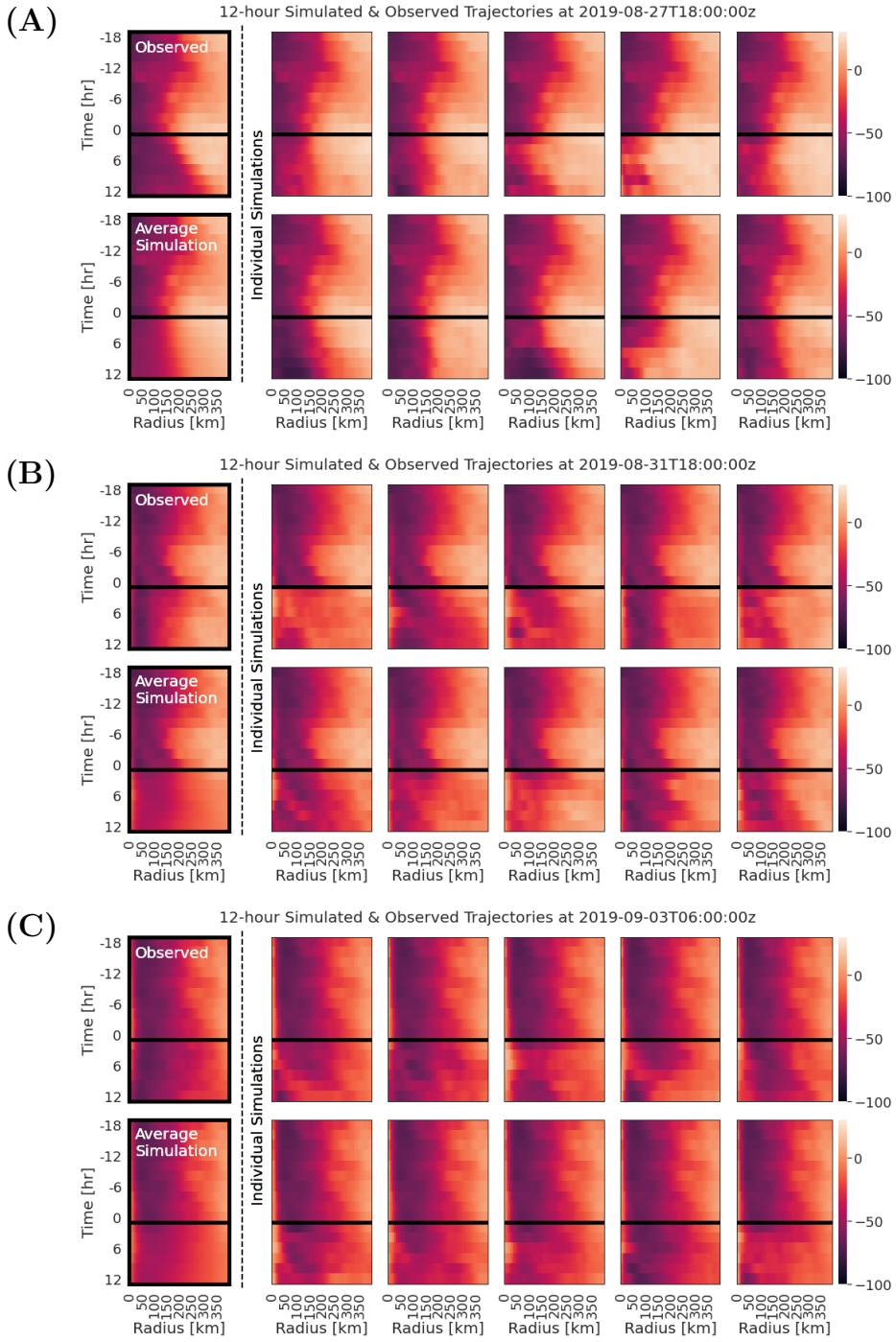


FIG. C9. **Hurricane Dorian [2019] 12-hr Structural Forecasts:** The *observed* structural trajectory is shown in the top left corner of each row. To the right, we see 10 *individual simulations* of radial profiles (averaged over all quadrants) at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner, shows the *average simulation* over 64 simulated trajectories.

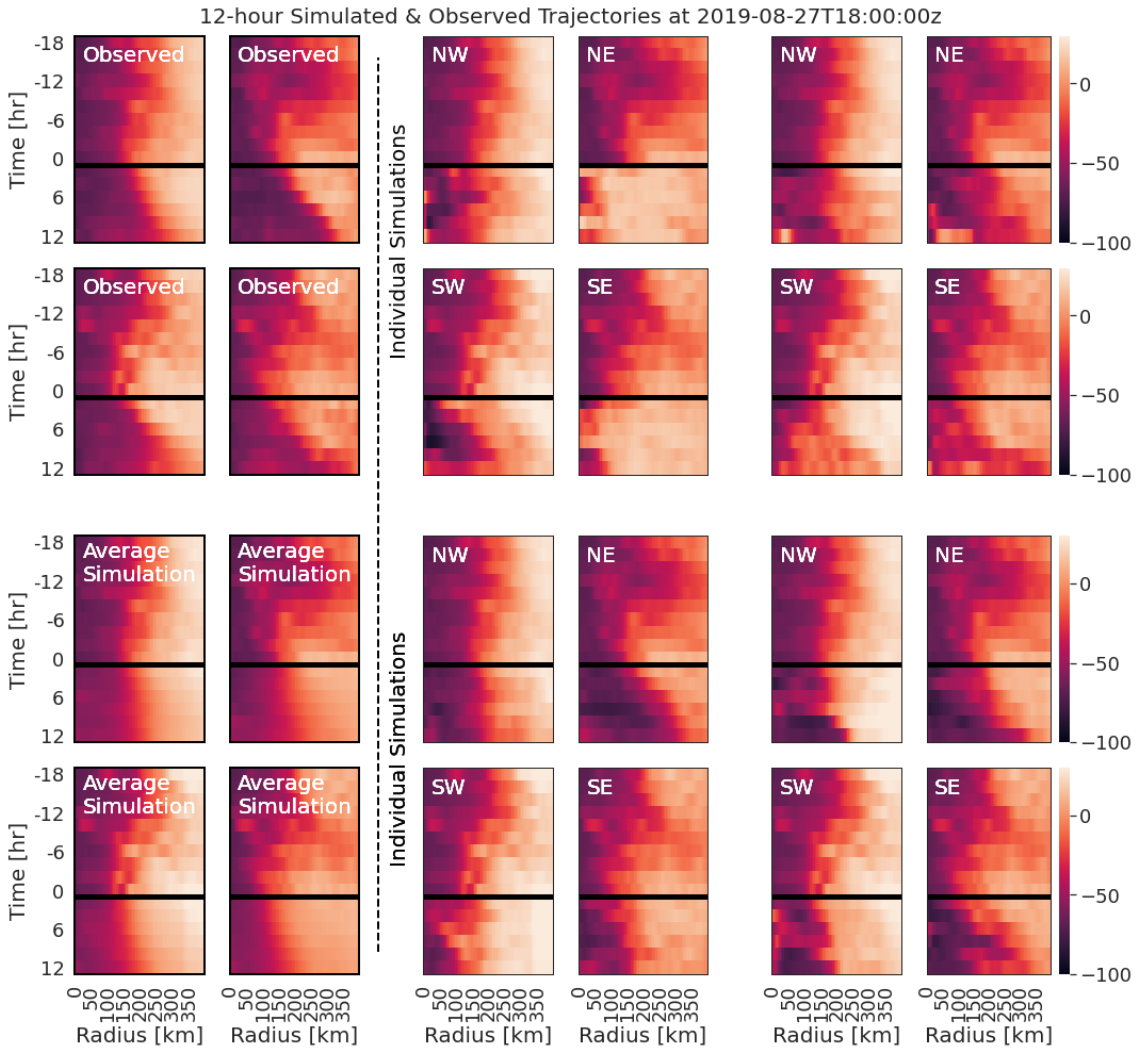


FIG. C10. **Hurricane Dorian [2019] Forecasts By Quadrant (A):** As Figure C9, but broken down by quadrant for example (A) only. The *observed* structural trajectory is shown in the top left corner. To the right, we see four *individual simulations* of radial profiles by quadrant at 12-hour lead times. All radial profiles above the black horizontal line are observed, while profiles below the black line are simulated. The bottom left corner shows the *average simulation* over 64 simulated trajectories.