# Predicting Tropical Cyclone Track Forecast Errors Using a Probabilistic Neural Network

M. A. FERNANDEZ,[a] ELIZABETH A. BARNES,[a] RANDAL J. BARNES,[b] MARK DEMARIA,[c] MARIE MCGRAW,[c] GALINA CHIROKOVA,[c] AND LIXIN LU[c]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*
[b] *Department of Civil, Environmental, and Geo-Engineering, University of Minnesota, Minneapolis, Minnesota*
[c] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*

(Manuscript received 25 July 2024, in final form 17 February 2025, accepted 4 March 2025)

ABSTRACT: A new method for estimating tropical cyclone track uncertainty is presented and tested. This method uses a neural network to predict a bivariate normal distribution, which serves as an estimate for track uncertainty. We train the network and make predictions on forecasts from the National Hurricane Center (NHC), which currently uses static error distributions based on forecasts from the past 5 years for most applications. The neural network–based method produces uncertainty estimates that are dynamic and probabilistic. Further, the neural network–based method allows for probabilistic statements about tropical cyclone trajectories, including landfall probability, which we highlight. We show that our predictions are well calibrated using multiple metrics, that our method produces better uncertainty estimates than current NHC approaches, and that our method achieves similar performance to the Global Ensemble Forecast System. Once trained, the computational cost of predictions using this method is negligible, making it a strong candidate to improve the NHC's operational estimations of tropical cyclone track uncertainty.

SIGNIFICANCE STATEMENT: Tropical cyclones affect millions of people across the planet, and accurate uncertainty estimates for their trajectories are vital for informing risk, evacuations, and mitigation planning. For most applications, the National Hurricane Center currently quantifies uncertainty using a historical-based estimate that remains static for the entire season. We propose a method that uses machine learning to dynamically estimate track uncertainty using inputs that are specific to the storm being forecast. Our method produces a probability distribution, specifically a bivariate normal, which presents decision-makers and researchers with a more informative assessment of tropical cyclone track uncertainty. We demonstrate that our method has many appealing properties, including the ability to produce landfall probabilities and outperform currently used National Hurricane Center methods.

KEYWORDS: Tropical cyclones; Probability forecasts/models/distribution; Numerical weather prediction/forecasting; Machine learning; Neural networks

## 1. Introduction

Tropical cyclones (TCs) expose populations and assets to risk around the world, with negative effects on population well-being (Berlemann and Eurich 2021). Forecasting centers have improved TC track accuracy through better modeling and techniques (Heming et al. 2019), though TC track uncertainty has not undergone similar improvements (Dunion et al. 2023). Uncertainty quantification is particularly important for TC forecasting, informing risk assessment and disaster planning (including mitigation and evacuations) on times scales from hours to days, as well as policy decisions when aggregated over entire TC seasons. Here, we focus on estimating TC track uncertainty directly rather than making predictions of TC tracks with uncertainty as a by-product.

The National Weather Service (NWS) National Hurricane Center (NHC) has long recognized the need to provide uncertainty information with its deterministic TC forecasts. The first NHC track uncertainty product was the strike probabilities, which became operational in 1983 (Sheets 1985). The strike probabilities only considered track uncertainty and were replaced by wind speed probabilities (WSPs) in 2006, which take into account the uncertainty in the track, intensity, and wind structure forecasts (DeMaria et al. 2009). The NHC also provides the graphical "cone of uncertainty," which shows the area enclosed within the 67th percentile of the NHC's historical track error distributions (for a given year, the historical is the previous 5 years). In addition, the NHC's storm surge watches and warnings implemented in 2017 are based on a probabilistic storm surge (P-surge) model (Penny et al. 2023), which uses an ensemble of statistically generated track and intensity forecasts to drive a simplified surge model.

Although uncertainty information is included in many NHC products, the underlying probabilities are determined almost entirely from historical error distributions and include little information about the specific forecast situation. For example, the variability in the wind forcing for the P-surge model (Penny

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-24-0066.s1.

*Corresponding author*: M. A. Fernandez, mafern@colostate.edu

DOI: 10.1175/AIES-D-24-0066.1 e240066

et al. 2023) and the track and intensity uncertainty for the WSP model are based on historical forecast errors from the NHC, the Central Pacific Hurricane Center (CPHC), or the Joint Typhoon Warning Center (JTWC) from the past 5 years. A method to add situational dependence to the track uncertainty was added to the WSP model in 2011 by stratifying the NHC track errors by the Goerss predicted consensus error (GPCE; Goerss 2007). GPCE uses linear regression to predict the track error of a consensus model based on the spread of the models in the consensus and the TC intensity. However, the wind structure variability in the WSP model is still determined from historical error distributions, and the GPCE input only has a small impact on the track error distributions (DeMaria et al. 2013).

Another weakness of current operational track forecast uncertainty estimates is that the error estimates are circular. For example, the NHC cone of uncertainty uses static radii for each forecast basin (Atlantic, eastern Pacific, central Pacific) and forecast time so that for each point along the forecast track, the error estimate is a circle. The GPCE product estimates the expected error of the consensus forecast and then scales that to a radius that includes the forecast track ~68% of the time, similar to that used in the cone. Hansen et al. (2011) developed a generalized version of GPCE called GPCE along–across (GPCE-AX) that includes separate regression equations for the along- and across-track errors so that the uncertainty areas are not circular. However, GPCE-AX is not used operationally by the NHC or CPHC in any of their public-facing forecast uncertainty products.

A probabilistic method that has been explored previously is the use of ensemble forecasts for estimating TC track and track uncertainty (Dupont et al. 2011; Bonnardot et al. 2019; Kawabata and Yamaguchi 2020; Zhang and Yu 2017; Dunion et al. 2023; Wilks et al. 2009). Ensemble systems, such as those based on the Global Forecast System (GFS) and European Centre for Medium-Range Weather Forecasts (ECMWF) global models, support forecasters in understanding possible track scenarios when making their deterministic track forecasts. An example of such an ensemble system is the Global Ensemble Forecast System (GEFS; Zhou et al. 2017; Guan et al. 2022), which is based on the GFS. While ensemble systems provide useful information, the public-facing probabilistic products from operational centers need to be consistent with their deterministic forecasts. For example, if the NHC track forecast shows a landfall in Miami, but all or most of the ensemble members are north of that position, the contradiction between the products could cause considerable confusion. Therefore, corrections to ensemble systems are needed if they are used for public-facing uncertainty products.

Inclusion of situationally dependent forecast uncertainty in NWS products remains a high priority. An emerging method for estimating uncertainty is through the use of machine learning methods (Haynes et al. 2023; Barnes and Barnes 2021; Foster et al. 2021; Guillaumin and Zanna 2021; Gordon and Barnes 2022). Recent work by Barnes et al. (2023) used an artificial neural network to predict the parameters of a probability distribution as a means of quantifying uncertainty (Nix and Weigend 1994a,b) for TC intensity forecasting, with applications to rapid intensification prediction. Here, we ask whether we can make meaningful predictions of TC track uncertainty for specific TCs in a well-calibrated probabilistic framework.

To answer that question, we use a similar framework to Barnes et al. (2023). We task a neural network with predicting the parameters of a distribution (in this case, a bivariate normal distribution) that estimates TC track latitude and longitude uncertainty in kilometers. Specifically, our framework is designed for use in NHC operations, i.e., we have trained and tested our method on official NHC forecasts so that the uncertainty products will maintain consistency with the NHC official forecast.

There are many benefits to quantifying uncertainty using the approach detailed in this work. Like the historical-based measures of uncertainty, our predictions are data-driven: in this case, we use a neural network. Unlike the current operational methods, our bivariate normal predictions are based on forecast-specific inputs, including environmental variables and dynamical model outputs, and can vary through the correlation and two variance parameters. The use of a defined distribution means our method does not require running expensive ensembles, or calculating statistics from a limited population of ensemble members, but can use output from those systems as input to the network. Because of the probabilistic approach and forecast-specific inputs, the predictions returned by this network are a plausible alternative to the historically derived track uncertainty estimates used in NHC and CPHC operations.

## 2. Prediction framework

Our goal is to make well-calibrated probabilistic predictions of TC track uncertainty using forecast-specific inputs. We accomplish this task using a neural network that predicts the five parameters of a bivariate normal distribution, which serves as our estimated TC track uncertainty. The bivariate normal distribution was also used for the original strike probability product. Other choices for the distribution were explored; however, the bivariate normal was effective and simplicity won out over other potential choices. The dataset we use includes forecast-specific inputs and true errors from NHC and CPHC forecasts covering seasons from 2013 through 2023, a period chosen to balance model availability and quality with sample size. In the remainder of the discussion, the term NHC is assumed to include the NHC and CPHC forecasts.

### a. *Neural network*

The full prediction framework is shown schematically in Fig. 1. A set of forecast-specific, model-based, and environmental variables is first normalized by removing the mean and dividing by the standard deviation before being passed to the dense layers of the network. A detailed list and description of these inputs are shown in Table 2. There are two fully

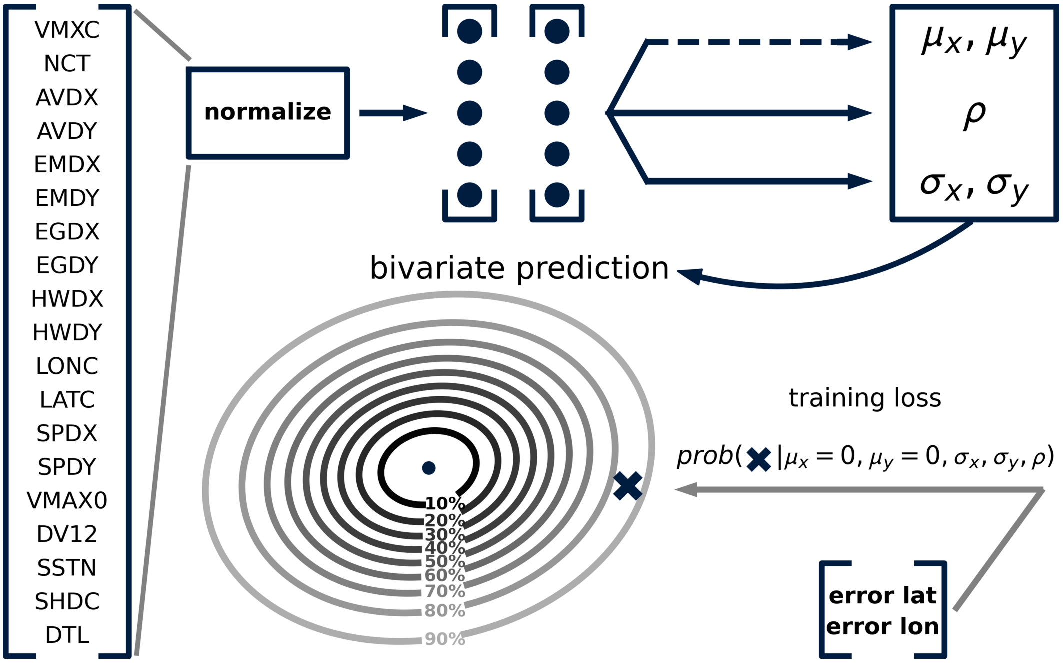

FIG. 1. Schematic showing the network architecture and format of model predictions. In many of the following figures, these predictions are used to construct a two-dimensional CDF, defined by the Mahalanobis distance (Mahalanobis 1936). Each labeled ellipse encloses the integrated probability out to that distance. Larger percentiles enclose more of the probability; thus, confidence that the truth falls within a percentile increases with percentile. Inputs are described in Table 2.

connected dense layers with five nodes each, both of which use the rectified linear unit (ReLU) activation function.

The outputs, which are the five parameters of a bivariate normal distribution in two dimensions (subscripts $x$ and $y$ represent longitude and latitude), are then each handled separately. The standard deviations ($\sigma_x$, $\sigma_y$) are passed through a softplus layer, which has the form $\ln(1 + e^{\alpha})$ and shifts the range to $[0, \infty]$. The correlation $\rho$ is passed through a hyperbolic tangent layer tanh, which shifts the range to $[-1, 1]$. There is no range restriction on the means ($\mu_x$, $\mu_y$). Nonzero $\mu_x$ and $\mu_y$ act as corrections to the NHC forecast for TC latitude and longitude. Our priority is to produce meaningful estimates for track uncertainty, so in all of our results, we freeze $\mu_x$ and $\mu_y$ to zero. Testing without this restriction showed that the predicted $\mu_x$ and $\mu_y$ are small. This is not surprising since the track bias of the NHC forecasts is generally much smaller than the mean track error for large samples. All of the parameters are then rescaled to return to their original units immediately prior to output.

The network trains by minimizing the loss defined by the negative log probability of the NHC forecast error (i.e., the truth, as shown in the lower-right bracketed list in Fig. 1), given the predicted bivariate normal distribution. Note that in Fig. 1, the example bivariate normal shown is the cumulative distribution function (CDF), rather than the probability density function (PDF) from which the loss is calculated. The loss function penalizes narrow predictions when the true forecast error is large (i.e., the truth lies outside the bulk of the distribution) and penalizes broad predictions when the true forecast error is small (i.e., the entire distribution is relatively flat). Early stopping is used for the training with a patience of 250 epochs. The batch size is 64 with a learning rate of 0.0001.

The data are separated into the Atlantic and eastern/central Pacific basins. There are less than 2000 central Pacific samples in the dataset, which is too small to reasonably train a separate network on. We trained networks without including the central Pacific samples and found only marginal changes to the eastern Pacific estimates; thus, we combined these basins. The data are further separated into lead times every 12 h up to 5 days. A separate network is trained for each basin and lead time combination. We tested the effectiveness of predicting all lead times using a single network, where lead time was used as an input feature. Predictions made using this setup generally evolved more smoothly over lead time but tended toward more circular predictions (i.e., the correlation parameter $\rho$ was consistently near zero) and did not noticeably improve or degrade the predictions overall (not shown).

For each network, the data are split into training, validation, and testing sets. The testing set, which is all samples from a given year, is split off first. The validation set is 200 randomly selected samples from the remaining data, and the training set is the rest of the samples. For the results shown in this work, we use leave-one-year-out method, which iterates through all potential years for the testing set. Thus, the total number of trained networks is 220 (two basins, 10 lead times, 11 years). In this way, we make predictions for all forecasts over the entire dataset without ever using the testing samples for training.

TABLE 1.. Short name and description of the label variables (i.e., the truth) used in training our networks.

| Label variable | Description |
|---|---|
| OFDX (km) | Distance east of the best track position from the NHC official forecast |
| OFDY (km) | Distance north of the best track position from the NHC official forecast |

While our predictions are flexible and dynamic, the method of producing those predictions (neural networks) is opaque. In section S4 in the online supplemental material, we use the explainable artificial intelligence (XAI) method Shapley additive explanations (SHAPs; Lundberg and Lee 2017) to explore feature relevance, i.e., how our network arrived at its predictions.

### b. Dataset

The dataset used for labels (truth) and inputs is listed in Tables 1 and 2. Each of these variables is recorded for official forecasts made by the NHC during the 2013–23 seasons, with potential lead times from 12 to 120 h. The NHC currently does not make forecasts for 84- and 108-h lead times and did not make 60-h lead time forecasts until 2019. NHC track points at these times were obtained by linear interpolation if an NHC forecast was available before and after each of those times. The forecast time in all cases is based on synoptic time (i.e., 0000, 0600, 1200, 1800 h UTC). The dataset includes 186 TCs in the Atlantic and 217 TCs in the eastern/central Pacific for a total of over 40 000 forecasts. This is an updated version (now including the 2022 and 2023 seasons) of the raw dataset used in Barnes et al. (2023) to predict TC intensity.

Labels (i.e., the truth) are derived from the best track verification, a poststorm analysis that includes the TC track among other TC characteristics (Landsea and Franklin 2013). The labels are the distance, in kilometers, between the best track latitude and longitude and the relevant forecasted latitude and longitude, i.e., the error between forecast and true TC location.

Inputs include both dynamical model forecasts and TC predictors from statistical models. The TC predictors were included because the performance of the dynamical models has some dependence on these. For example, one of the most significant predictors of track error in GPCE is the TC intensity.

The dynamical models were chosen based on their track forecast skill and the availability of a long data record for training. Based on these criteria, one regional hurricane model and three global models were included as follows: Hurricane Weather Research and Forecasting Model (HWRF) (Tallapragada 2016), Met Office (UKMet) global model (Bush et al. 2023), GFS (Zhou et al. 2019), and ECMWF (Magnusson et al. 2021). Outputs from these models are not available until after the NHC official forecast is issued; thus, an interpolated version (based on the previous forecast cycle) is used as input. The interpolated models are referred to as “early” models. We use the early models to be consistent with what is available to NHC forecasters at advisory time (Cangialosi et al. 2023) and so the uncertainty estimates can be determined shortly after the advisories are issued.

The average of the four early track model forecasts is called the “consensus” and can be calculated as long as at least two of the four input models are available at a given forecast time. The neural network inputs from the early models are the deviations from the consensus forecast (from AVDX through HWDY in Table 2). For missing models, the track forecast is replaced by the consensus of the available models, so the deviations are zero for that model. The number of models (NCT) is also included as a predictor because the track errors might

TABLE 2. Input variables used in training our networks.

| Input variable | Description |
|---|---|
| VMXC [kt; (1 kt ≈ 0.51 m $s^{-1}$)] | Max wind of the consensus forecast |
| NCT (No.) | Number of models included in the consensus forecast |
| AVDX (km) | Distance east of the early GFS forecast from the consensus forecast |
| AVDY (km) | Distance north of the early GFS forecast from the consensus forecast |
| EMDX (km) | Distance east of the early ECMWF forecast from the consensus forecast |
| EMDY (km) | Distance north of the early ECMWF forecast from the consensus forecast |
| EGDX (km) | Distance east of the early UKMet forecast from the consensus forecast |
| EGDY (km) | Distance north of the early UKMet forecast from the consensus forecast |
| HWDX (km) | Distance east of the early HWRF forecast from the consensus forecast |
| HWDY (km) | Distance north of the early HWRF forecast from the consensus forecast |
| LONC (°E) | Longitude of the consensus forecast |
| LATC (°N) | Latitude of the consensus forecast |
| SPDX (kt) | Average eastward speed from DSHP in the 24 h preceding the forecast |
| SPDY (kt) | As in SPDX, but for the northward speed of the TC |
| VMAX0 (kt) | Max wind at the start of the forecast |
| DV12 (kt) | Intensity change in the 12 h preceding the forecast |
| SSTN (°C) | Average SST in the 24 h preceding the forecast |
| SHDC (kt) | Average 850–200-hPa vertical shear in the 24 h preceding the forecast |
| DTL (km) | Distance to the nearest major landmass at forecast time |

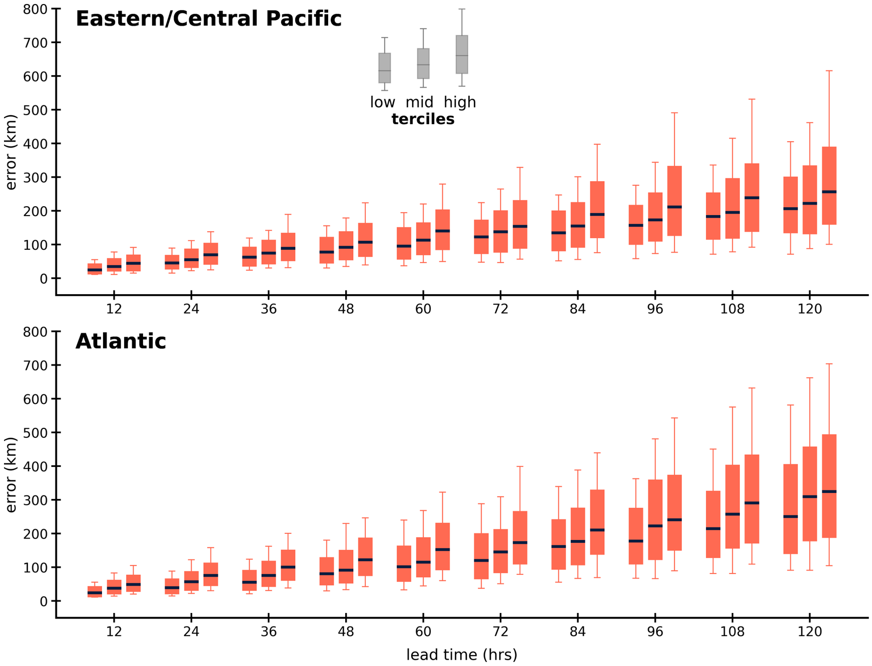

FIG. 2. IQR vs error. Boxplots show the distribution of forecast error (filled area spans the 25th–75th percentile, whiskers out to the 10th and 90th percentile) associated with the lower, middle, and upper terciles of IQR for each lead time. The IQR is a measure of the width of the predicted bivariate distribution.

be larger when some of the skillful models are not available. Most of the dataset has all four models available (72%), with a small percentage having fewer than two models (2.5%).

The TC predictors include three basic TC parameters (latitude and longitude of the TC center and the maximum wind). The latitude and longitude are from the consensus forecast (LATC and LONC), and the maximum wind (VMXC) is from a consensus of four skillful early intensity models comprised of the GFS, HWRF, and two statistical–dynamical intensity models (Barnes et al. 2023).

Seven additional TC predictors are obtained from the statistical–dynamical Decay-Statistical Hurricane Intensity Prediction Scheme (D-SHIPS) (DeMaria et al. 2022). These are comprised of the 0-h maximum wind (VMAX0, sustained 1-min average estimate at synoptic time), the change in maximum wind over the 12-h period ending at the start of the forecast (DV12), the eastward and northward components of the TC translational velocity (SPDX, SPDY), the distance of the TC center from major landmasses (DTL), the sea surface temperature (SSTN), and the 850–200-hPa wind shear averaged from 0 to 500 km (SHDC). The last six of the above predictors require a track forecast, which is obtained from an interpolated (early) version of the NHC official forecast from the previous cycle in the D-SHIPS model, which is often run prior to the official TC genesis declaration.

Analyses and figures presented in this work use our estimates of the uncertainty of the NHC official forecast, which could be used as input for other hazard products such as NHC's wind speed probability or P-surge models. However, predictions can also be made with respect to the consensus forecast, which would be available before the NHC forecast is issued due to the use of early model input. The consensus uncertainty could be used as guidance by NHC forecasters for their official forecasts and products such as the tropical cyclone discussion, which sometimes include qualitative descriptions of forecast confidence.[1]

### c. *Model calibration*

Many metrics support determining the calibration and validity of probabilistic models (Gneiting and Raftery 2007), and here, we showcase two such metrics: the interquartile range (IQR) versus error and the probability integral transform (PIT; Dawid 1984). Figure 2 shows the IQR versus true error (the labels used in training) for both the eastern Pacific and the Atlantic basins.

IQR values are computed as the difference between the 75th and 25th percentiles for each predicted bivariate normal and are thus a measure of the width of each predicted distribution. In Fig. 2, the IQR is divided in three bins for each lead time (lead time indicated along the horizontal axis): the lower, middle, and upper terciles of IQR for the set of predictions for that basin and lead time. The true errors associated with each of these bins are shown, with the median (solid line), the 25th–75th percentile (filled), and 10th and 90th percentile (whiskers) all indicated. For well-calibrated networks,

[1] https://www.nhc.noaa.gov/aboutnhcprod.shtml.

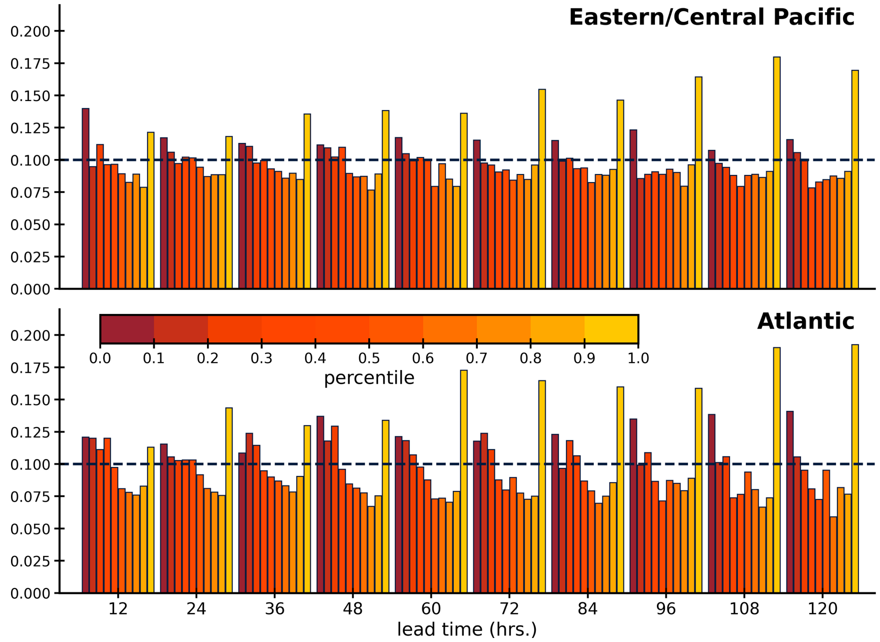

FIG. 3. PIT histogram for the (top) eastern Pacific and (bottom) Atlantic for all forecast lead times. This metric describes how often the truth falls into each decile of the predictions (10th, 20th, 30th, etc.). A perfectly calibrated probabilistic model would have a uniform distribution of 0.1.

we expect the median error and the error range to be larger for larger IQR. This is evident for each lead time in Fig. 2, where the distribution shifts to higher error as we move from the lowest IQR tercile up to the highest IQR tercile. Static error distribution parameters such as those used in the cone of uncertainty are not able to capture this variability other than the basin and lead time dependence.

PIT histograms, shown in Fig. 3, quantify how often the truth falls into a certain percentile of the predicted bivariate normal distribution's CDF. PIT values are shown for every 10% increment, e.g., the leftmost bar for each lead time shows the fraction of the time the truth falls between the 0th and 10th percentile, the next bar is for the 10th–20th percentile, and so on. A perfectly calibrated model would be uniform with a constant value of 0.1, indicated in the figure by a dashed horizontal line. For the eastern/central Pacific, our networks are making too many wide and narrow predictions (the rightmost and leftmost bars for each lead time are larger than 0.1). The Atlantic shows the same, but slightly stronger, bias as the eastern Pacific. However, the values for most of the other bins are not too far from 0.1.

One way of quantifying how well calibrated our predictions are is to compare the PIT-D statistic, which measures the deviation of our PIT histogram from a uniform distribution, to the expected deviation. The PIT-D statistic is given by $D = \sqrt{1/B\sum_k(b_k - 1/B)}$, while the expected deviation is given by $E[D] = \sqrt{(1 - 1/B)/(T \times B)}$, where $B$ is the number of bins, $T$ is the number of samples, and $k$ indicates the summation over each bin (Nipen and Stull 2011; Bourdin et al. 2014). Our predictions range from $D = 0.011$ up to $D = 0.037$, while the expected deviation is between $E[D] = 0.005$ and $E[D] = 0.010$.

## 3. Results

Satisfied that our framework produces reasonable and well-calibrated uncertainty predictions, we turn to the use of these predictions. In particular, we analyze our predictions for all forecasts made by the NHC from 2013 through 2023 in the eastern/central Pacific and Atlantic basins. We do this by using a leave-one-year-out method: we train our network on all but 1 year and then predict that left-out year. We iterate through each left-out year to obtain predictions for every forecast without the network's seeing that year in its training.

Figure 4 shows two examples of forecasts with our predicted bivariate normal CDF overlaid in the red-to-yellow shading. The top panels show the forecasts made for Hurricane Nicole at 0600 UTC 5 October 2016, and the bottom panels show the forecasts made for Hurricane Matthew at 0600 UTC 4 October 2016. In both cases, forecasts are shown out to 5 days. The bivariate normal CDF is centered at the NHC official forecast location; we do not fit the location parameters ($\mu_x$ and $\mu_y$) of our bivariate normal, as described in section 2. We show two concurrent storms in the same basin to emphasize that our method predicts uncertainties based on forecast-specific inputs. The larger predicted uncertainties for Hurricane Nicole reflect that it was difficult to forecast. Hurricane Matthew was easier to forecast with smaller error, also reflected in our predicted uncertainties.

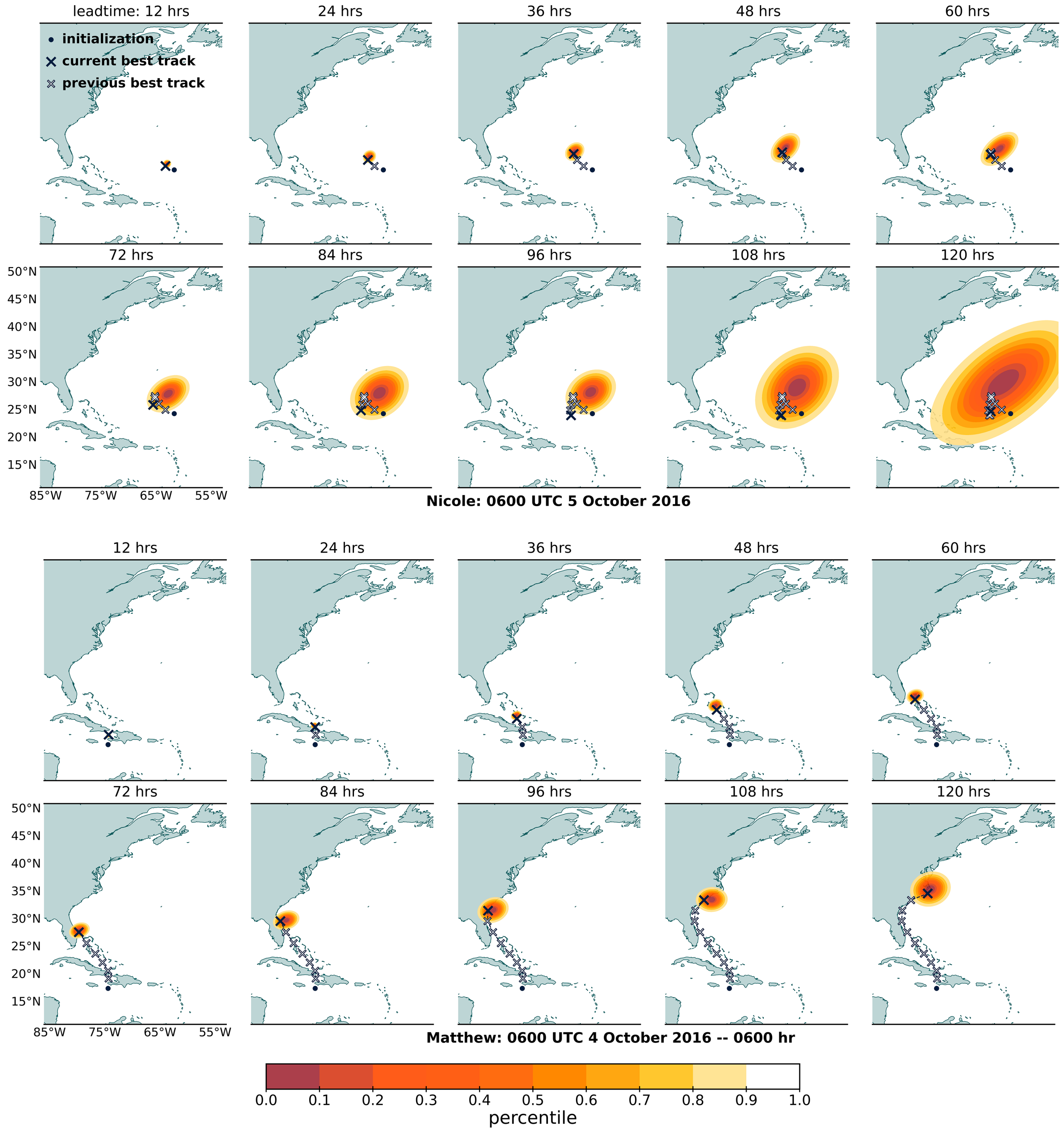

FIG. 4. Two examples of our predictions, as described in Fig. 1. The initialization time is fixed, and the forecast every 12 h out to 5 days is shown, along with the best track reconstruction.

Among the forecasts, there are several examples that highlight the usefulness of the correlation parameter (i.e., the flexibility of our predicted bivariate normal shape). This is especially apparent for Nicole, where the forecast was consistently to the northeast of the truth. Our predicted bivariates point in a northeast–southwest direction, emphasizing that the uncertainty is larger along that axis.

### a. *Comparison with NHC cone, GPCE radii, and GEFS*

Comparing our probabilistic estimate of track uncertainty directly to the NHC cone or GPCE radii is difficult; the cone and GPCE only provide a single radii value at each forecast time, but our method estimates the full error distribution. However, one metric that can be used is the continuous ranked probability score (CRPS) (Gneiting and Raftery 2007). The

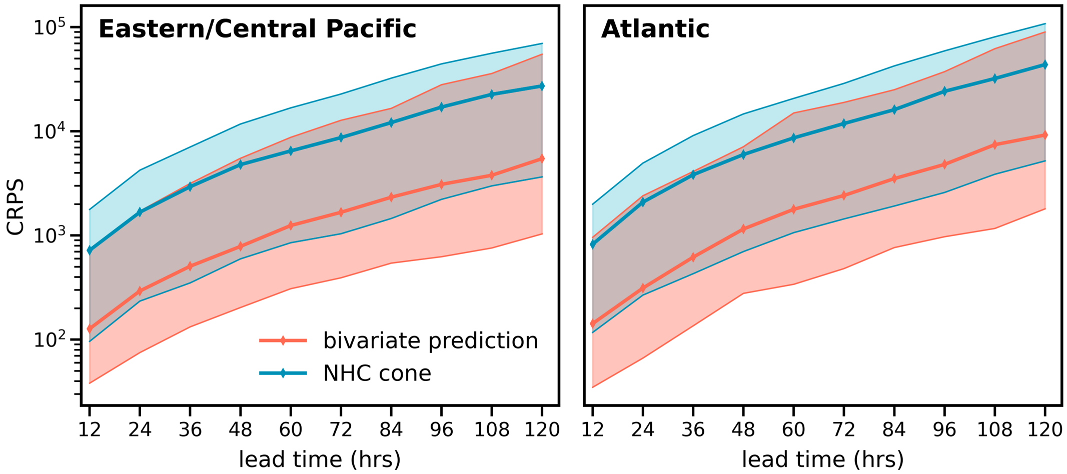

FIG. 5. CRPS for the NHC cone and the bivariate predictions as a function of lead time. The median for each is shown as the solid lines, while the shaded area encloses the 10th–90th percentile of CRPS values. A lower value of CRPS indicates a better prediction (a CRPS value of zero indicates a perfect prediction, with the entirety of the prediction weight at the truth, e.g., a delta function).

CRPS collapses to the mean absolute error when both parts are deterministic (both the prediction and truth CDFs are step functions), so it can be thought of as an extension to the mean absolute error that allows for a probabilistic component.

For our two-dimensional case, we calculate the one-dimensional CRPS along both the latitude and longitude and multiply these, only integrating over the quadrant in which the truth falls. For the NHC cone or GPCE, which are symmetric (circular), no further adjustments are necessary. To account for the variable shape of bivariate normal predictions, the prediction CDF used to calculate the CRPS is the distribution conditioned on the truth along the other axis; e.g., to calculate the CRPS along the latitude axis, we condition on the true longitude error. This is further explained in section S1.

Figure 5 shows the results of the CRPS calculation for our predictions (red) and the NHC cone (blue). As with mean absolute error, a lower CRPS value is better. The CRPS is calculated for all forecasts, and the median for these is shown as a solid line, with the 10th–90th percentiles shaded. According to the CRPS metric, our predictions are a better estimate of the true error than the NHC cone for a majority of the forecasts. This is unsurprising, as the NHC cone is static throughout a season and has a fixed circular symmetry.

We calculated the CRPS for the GPCE predictions as well and found these to be very similar to the NHC cone. Using the standard deviation of the GEFS member displacements from the GEFS mean, and the corresponding correlation, we construct bivariate normal predictions and calculate the CRPS for GEFS. We find the GEFS CRPS to be very similar to our bivariate predictions. We reiterate that running an ensemble such as GEFS has a much higher computational cost than the method presented here. Both GPCE and GEFS CRPS are presented in section S2.

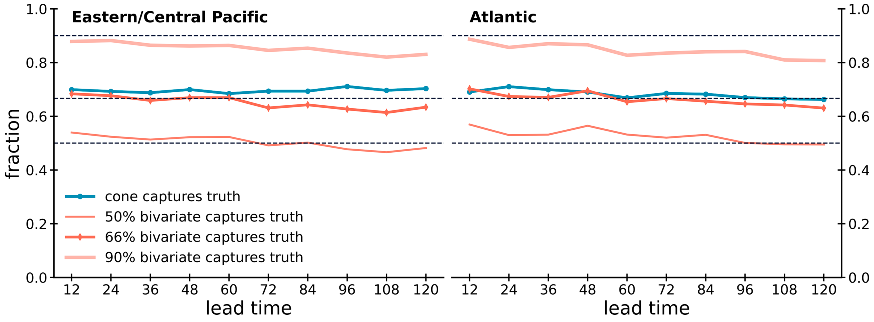

FIG. 6. Fraction of cases where the bivariate predictions (red) and the NHC cone (blue) capture the truth as a function of lead time. Shown are several percentiles for the bivariate; from thinnest to thickest, the 50th, 66th, and 90th percentile ellipses, respectively. This emphasizes the strength of using the distribution for prediction and allows for both probabilistic and tailored deterministic predictions (e.g., minimizing misses or minimizing false alarms).

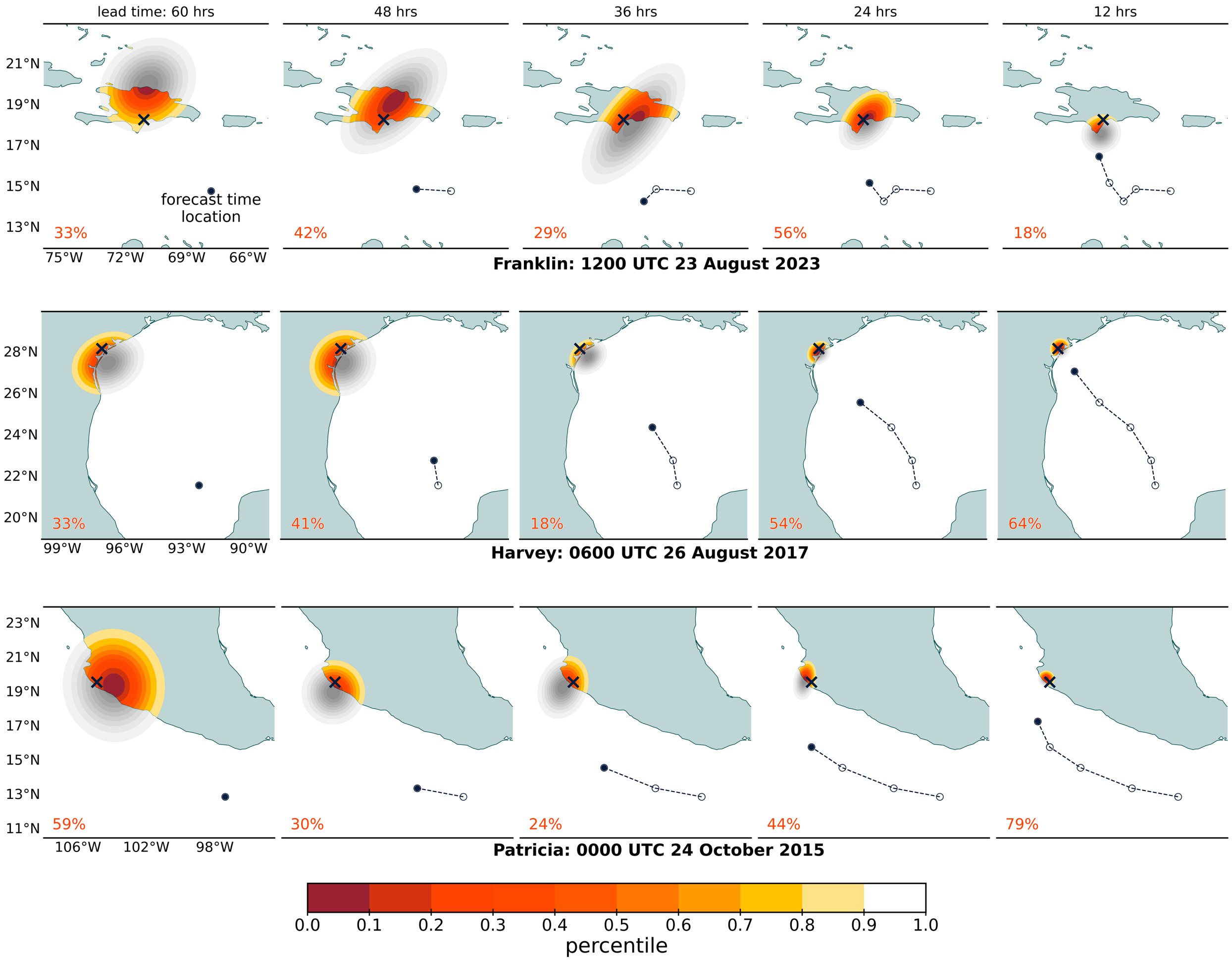

FIG. 7. Three examples of predictions for TCs as they make landfall. For each storm, the nearest synoptic time to landfall, and the corresponding location, is used. The forecasted time is held fixed (panels go from longer to shorter lead time). The landfall location is indicated by a cross, and the forecast initialization location is indicated by a solid dot (unfilled dots for previous times). Predictions are shaded red-to-yellow over land and in grayscale over the ocean. The distribution integrated over land gives a probability for the TC to make landfall. The probability (as a percent chance) is shown in the lower-left corner of each panel.

We can also make a comparison by choosing a specific percentile of our distribution and comparing only the associated ellipse, though this undermines one of the main strengths of our predictions, to the NHC cone or GPCE. With this inhibited version of our prediction, we can look at the binary question of whether each of the predictions captures the truth.

Figure 6 shows the fraction of forecasts where the NHC cone captured the true TC location (blue) and the fraction of forecasts where several of our percentile ellipses (red) captured the truth. We can use any percentile from our predictions, but we show only three: the 50th, 66th, and 90th percentile ellipse capture fractions. The dashed lines show perfect calibration; e.g., the 50th percentile of our distributions captures the truth 50% of the time. The 66th percentile ellipse in Fig. 6 remains remarkably close to the perfect 66% dashed line. This supports the results in Fig. 3 but additionally emphasizes the flexibility of our method. Specifically, our method allows for a subjective choice of either minimizing misses (using a higher percentile ellipse or setting a higher percentile threshold) or minimizing false alarms (using a lower percentile ellipse or a lower percentile threshold).

## b. *Landfall events*

Using our method, we can make probabilistic statements about landfall by integrating the portion of our predicted uncertainty that is over land at each forecast to obtain a probability of the TC making landfall at that time. Several examples of this are shown in Fig. 7. The top panel shows Hurricane Franklin (2023) making landfall over the Dominican Republic, the middle panel shows Hurricane Harvey (2017) making landfall over Texas in the United States, and the bottom panel shows Hurricane Patricia (2015) making landfall over Jalisco in Mexico. As expected, the uncertainty decreases as we approach the forecasted time.

We divide all forecasts into cases where the TC did make landfall and cases where it did not. Across all forecasts, there

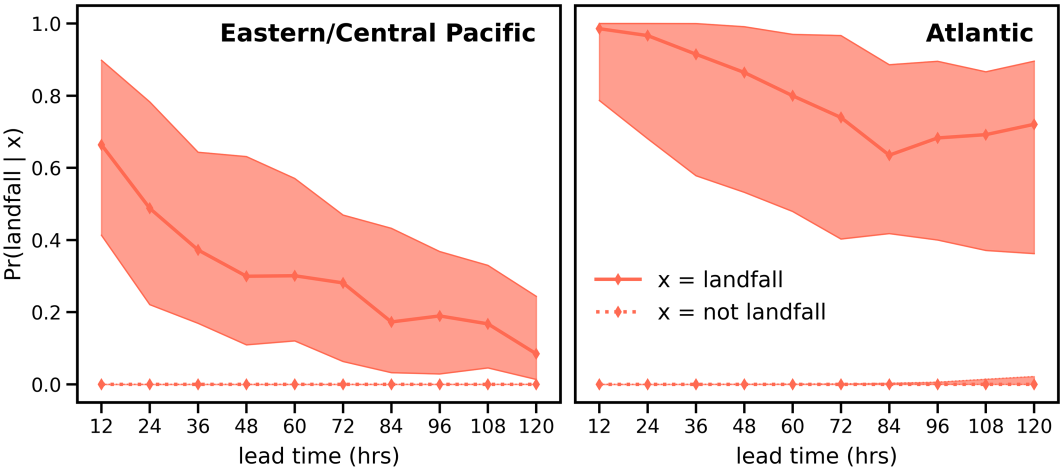

FIG. 8. Probability of landfall. The panels show the probability of landfall (bivariate integrated over land, see Fig. 7) for TCs in the (left) eastern/central Pacific and (right) Atlantic as a function of lead time. The solid lines show the median probability for cases when the TC did make landfall, prob(landfall | landfall), and the dotted lines show the median probability for cases when the TC did not make landfall, prob(landfall | not landfall). The shaded areas indicate the 25th–75th percentile of the distribution of landfall probabilities.

were 1889 instances where a TC made landfall in the Atlantic and 478 instances where a TC made landfall in the eastern/central Pacific. We checked the calibration metrics (PIT, IQR vs error) for this subsample of predictions and found that the predictions remained well calibrated (not shown).

Figure 8 shows the probability of landfall for both cases: when there was landfall *P*(landfall | landfall) and when landfall did not occur *P*(landfall | nolandfall). While the probability of landfall for cases when it did not occur is very low for both basins, the probability of landfall for cases when there was landfall looks fairly different. In the Atlantic, our predicted landfall probability is high for all lead times, with the median remaining above the 0.5 line throughout. The eastern/central Pacific probabilities decay strongly with lead time, likely due to the very small sample size available for training, a result of fewer landmasses in the path of eastern/central Pacific TCs. For example, at 120 h, there are only 19 forecasts where a TC made landfall in the eastern/central Pacific.

We repeat the preceding analyses for early forecasts (capture fraction, CRPS, landfall probability) and find similar performance in all cases. These are shown in section S3. We apply the preceding analysis (CRPS) to Atlantic landfall cases to compare our bivariate predictions to the NHC cone. We find a large improvement over the NHC cone, with the mean NHC cone CRPS approximately twice as large (worse) than our bivariate predictions. This is shown in section S6.

## 4. Conclusions

We have developed and tested a method of estimating tropical cyclone track uncertainty. Using forecast-specific inputs and the true forecast error as the label, we train a neural network to predict the parameters of a bivariate normal distribution. The distribution serves as our estimate of the TC track uncertainty for that forecast. The network is trained on a dataset from the NHC and CPHC, which includes 11 years (2013–23) of forecasts, 10 lead times (12–120 h), and the Atlantic and combined eastern and central Pacific basins. The loss used in training the network is the negative log probability of the truth (the difference between the forecast location and the best track reconstruction), given the predicted distribution.

We have shown that predictions using our method are well calibrated using the probability integral transform (PIT) metric. We have also compared the interquartile range (IQR) of our predictions to the true forecast errors. According to the continuous ranked probability score (CRPS), our method produces better uncertainty estimates than the NHC cone and the GPCE track uncertainty estimates and is comparable to predictions from the Global Ensemble Forecast System (GEFS). The probabilistic nature of our predictions allows for a subjective, expert-based decision on whether to emphasize minimizing false alarms or minimizing misses. We have also shown that a probabilistic approach can be used to robustly estimate the probability of landfall events.

The move toward a probabilistic estimate of track uncertainty is already a priority for forecasting centers (Dunion et al. 2023; Conroy et al. 2023). Currently, the NHC (and many other operational forecast centers) estimates TC track forecast errors using historical errors of their operational forecasts from the previous 5 years. These are static (the same for the entire season, circularly symmetric) and deterministic (a single-valued uncertainty estimate). Our method produces uncertainty estimates that are dynamic (forecast-specific, variable shapes) and probabilistic.

Once trained, the computational cost of predictions using our method is negligible, potentially making it more appealing than costly ensemble methods that may or may not have enough members for the distribution to converge. The method is flexible so that new models can be included, provided an adequate training sample is available. For example, the HWRF model used in this study is being replaced in NHC forecasts by the Hurricane Analysis and Forecast System (HAFS) model (Hazelton et al. 2021), which will require a different set of track models to be used as input. In addition, parameters from ensemble forecast systems such as ensemble spread can be added as input to the neural network. Thus, our method is a strong candidate to improve operational track uncertainty estimates.

*Acknowledgments.* This work was supported by NOAA Grant NA22OAR4590525.

*Data availability statement.* Operational model outputs, environmental variables, and best track data were obtained from the National Hurricane Center. The code and dataset used in this work can be found at https://github.com/mafern/tcane_track and will be given a permanent DOI on Zenodo at the time of publication.

## REFERENCES

Barnes, E. A., and R. J. Barnes, 2021: Controlled abstention neural networks for identifying skillful predictions for regression problems. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002575, https://doi.org/10.1029/2021MS002575.

——, ——, and M. DeMaria, 2023: Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environ. Data Sci.*, **2**, e15, https://doi.org/10.1017/eds.2023.7.

Berlemann, M., and M. Eurich, 2021: Natural hazard risk and life satisfaction—Empirical evidence for hurricanes. *Ecol. Econ.*, **190**, 107194, https://doi.org/10.1016/j.ecolecon.2021.107194.

Bonnardot, F., H. Quetelard, G. Jumaux, M.-D. Leroux, and M. Bessafi, 2019: Probabilistic forecasts of tropical cyclone tracks and intensities in the southwest Indian Ocean basin. *Quart. J. Roy. Meteor. Soc.*, **145**, 675–686, https://doi.org/10.1002/qj.3459.

Bourdin, D. R., T. N. Nipen, and R. B. Stull, 2014: Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. *Water Resour. Res.*, **50**, 3108–3130, https://doi.org/10.1002/2014WR015462.

Bush, M., and Coauthors, 2023: The second Met Office Unified Model–Jules Regional Atmosphere and Land configuration, RAL2. *Geosci. Model Dev.*, **16**, 1713–1734, https://doi.org/10.5194/gmd-16-1713-2023.

Cangialosi, J., B. Reinhart, and J. Martinez, 2023: 2023 National Hurricane Center verification report. 81 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2023.pdf.

Conroy, A., and Coauthors, 2023: Track forecast: Operational capability and new techniques—Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10). *Trop. Cyclone Res. Rev.*, **12**, 64–80, https://doi.org/10.1016/j.tcrr.2023.05.002.

Dawid, A. P., 1984: Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Roy. Stat. Soc.*, **147A**, 278–292, https://doi.org/10.2307/2981683.

DeMaria, M., J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. DeMaria, 2009: A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting*, **24**, 1573–1591, https://doi.org/10.1175/2009WAF2222286.1.

——, and Coauthors, 2013: Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, **28**, 586–602, https://doi.org/10.1175/WAF-D-12-00116.1.

——, and Coauthors, 2022: The National Hurricane Center tropical cyclone model guidance suite. *Wea. Forecasting*, **37**, 2141–2159, https://doi.org/10.1175/WAF-D-22-0039.1.

Dunion, J. P., and Coauthors, 2023: Recommendations for improved tropical cyclone formation and position probabilistic forecast products. *Trop. Cyclone Res. Rev.*, **12**, 241–258, https://doi.org/10.1016/j.tcrr.2023.11.003.

Dupont, T., M. Plu, P. Caroff, and G. Faure, 2011: Verification of ensemble-based uncertainty circles around tropical cyclone track forecasts. *Wea. Forecasting*, **26**, 664–676, https://doi.org/10.1175/WAF-D-11-00007.1.

Foster, D., D. J. Gagne II, and D. B. Whitt, 2021: Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002474, https://doi.org/10.1029/2021MS002474.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Goerss, J. S., 2007: Prediction of consensus tropical cyclone track forecast error. *Mon. Wea. Rev.*, **135**, 1985–1993, https://doi.org/10.1175/MWR3390.1.

Gordon, E. M., and E. A. Barnes, 2022: Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophys. Res. Lett.*, **49**, e2022GL098635, https://doi.org/10.1029/2022GL098635.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Guillaumin, A. P., and L. Zanna, 2021: Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002534, https://doi.org/10.1029/2021MS002534.

Hansen, J. A., J. S. Goerss, and C. Sampson, 2011: GPCE-AX: An anisotropic extension to the Goerss predicted consensus error in tropical cyclone track forecasts. *Wea. Forecasting*, **26**, 416–422, https://doi.org/10.1175/2010WAF2222410.1.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, 220061, https://doi.org/10.1175/AIES-D-22-0061.1.

Hazelton, A., and Coauthors, 2021: 2019 Atlantic hurricane forecasts from the global-nested hurricane analysis and forecast system: Composite statistics and key events. *Wea. Forecasting*, **36**, 519–538, https://doi.org/10.1175/WAF-D-20-0044.1.

Heming, J. T., and Coauthors, 2019: Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. *Trop. Cyclone Res. Rev.*, **8**, 181–218, https://doi.org/10.1016/j.tcrr.2020.01.001.

Kawabata, Y., and M. Yamaguchi, 2020: Probability ellipse for tropical cyclone track forecasts with multiple ensembles. *J. Meteor. Soc. Japan*, **98**, 821–833, https://doi.org/10.2151/jmsj.2020-042.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Curran Associates, Inc., 4765–4774, https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Magnusson, L., and Coauthors, 2021: Tropical cyclone activities at ECMWF. ECMWF Tech. Memo. 888, 140 pp., https://www.ecmwf.int/node/20228.

Mahalanobis, P. C., 1936: On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, **2**, 49–55.

Nipen, T., and R. Stull, 2011: Calibrating probabilistic forecasts from an NWP ensemble. *Tellus*, **63A**, 858–875, https://doi.org/10.1111/j.1600-0870.2011.00535.x.

Nix, D., and A. Weigend, 1994a: Estimating the mean and variance of the target probability distribution. *Proc. 1994 IEEE Int. Conf. Neural Networks (ICNN'94)*, Orlando, FL, Institute of Electrical and Electronics Engineers, 55–60, https://doi.org/10.1109/ICNN.1994.374138.

——, and ——, 1994b: Learning local error bars for nonlinear regression. *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., MIT Press, https://papers.nips.cc/paper/1994/hash/061412e4a03c02f9902576ec55ebbe77-Abstract.html.

Penny, A. B., L. Alaka, A. A. Taylor, W. Booth, M. DeMaria, C. Fritz, and J. Rhome, 2023: Operational storm surge forecasting at the National Hurricane Center: The case for probabilistic guidance and the evaluation of improved storm size forecasts used to define the wind forcing. *Wea. Forecasting*, **38**, 2461–2479, https://doi.org/10.1175/WAF-D-22-0209.1.

Sheets, R. C., 1985: The National Weather Service hurricane probability program. *Bull. Amer. Meteor. Soc.*, **66**, 4–13, https://doi.org/10.1175/1520-0477(1985)066<0004:TNWSHP>2.0.CO;2.

Tallapragada, V., 2016: Overview of the NOAA/NCEP Operational Hurricane Weather Research and Forecast (HWRF) modelling system. *Advanced Numerical Modeling and Data Assimilation Techniques for Tropical Cyclone Prediction*, Springer Netherlands, 51–106, https://doi.org/10.5822/978-94-024-0896-6_3.

Wilks, D. S., C. J. Neumann, and M. B. Lawrence, 2009: Statistical extension of the National Hurricane Center 5-day forecasts. *Wea. Forecasting*, **24**, 1052–1063, https://doi.org/10.1175/2009WAF2222189.1.

Zhang, X., and H. Yu, 2017: A probabilistic tropical cyclone track forecast scheme based on the selective consensus of ensemble prediction systems. *Wea. Forecasting*, **32**, 2143–2157, https://doi.org/10.1175/WAF-D-17-0071.1.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the Next Generation Global Prediction System. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.