A Statistical Model to Predict the Extratropical Transition of Tropical Cyclones

MELANIE BIELI

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

ADAM H. SOBEL

*Department of Applied Physics and Applied Mathematics, Columbia University, New York, and
Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York*

SUZANA J. CAMARGO

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

MICHAEL K. TIPPETT

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

(Manuscript received 4 March 2019, in final form 2 December 2019)

ABSTRACT

This paper introduces a logistic regression model for the extratropical transition (ET) of tropical cyclones in the North Atlantic and the western North Pacific, using elastic net regularization to select predictors and estimate coefficients. Predictors are chosen from the 1979–2017 best track and reanalysis datasets, and verification is done against the tropical/extratropical labels in the best track data. In an independent test set, the model skillfully predicts ET at lead times up to 2 days, with latitude and sea surface temperature as its most important predictors. At a lead time of 24 h, it predicts ET with a Matthews correlation coefficient of 0.4 in the North Atlantic, and 0.6 in the western North Pacific. It identifies 80% of storms undergoing ET in the North Atlantic and 92% of those in the western North Pacific. In total, 90% of transition time errors are less than 24 h. Select examples of the model's performance on individual storms illustrate its strengths and weaknesses. Two versions of the model are presented: an “operational model” that may provide baseline guidance for operational forecasts and a “hazard model” that can be integrated into statistical TC risk models. As instantaneous diagnostics for tropical/extratropical status, both models' zero lead time predictions perform about as well as the widely used cyclone phase space (CPS) in the western North Pacific and better than the CPS in the North Atlantic, and predict the timings of the transitions better than CPS in both basins.

1. Motivation

Tropical cyclones (TCs) moving into midlatitude regions can transform into strong extratropical cyclones, changing their structures and the nature of the hazards to coastal populations and infrastructure. In this transformation, which is called extratropical transition (ET), TCs lose their radial symmetry and become cold-cored systems with fronts. Transitioning cyclones typically

grow larger and accelerate their forward motion, producing intense precipitation, strong winds, and large surface water waves (Jones et al. 2003; Evans et al. 2017). In the western North Pacific and the North Atlantic, 40%–50% of all TCs undergo ET (Hart and Evans 2001; Kitabatake 2011), posing a serious threat to coastal regions in East Asia, Japan, northeast North America, and western Europe.

Forecasts of the timing of ET as well as of the track and structural evolution of the transitioning storm are

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-19-0045.s1>.

Corresponding author: Melanie Bieli, melanie@caltech.edu

Publisher's Note: This article was revised on 27 February 2020 to correct a misspelling of the first author's name when originally published and to provide additional information in the Acknowledgments section.

DOI: 10.1175/WAF-D-19-0045.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

difficult for numerical weather prediction models (e.g., Buizza et al. 2005; Fogarty 2010; Keller et al. 2011; Evans et al. 2017; Keller et al. 2019), as these forecasts depend on the representation of both the TC and the midlatitude circulation into which the TC is moving. Errors in forecasts of ET can be large as a result of nonlinear interactions between errors in the simulation of the TC structure and of the large-scale environment (Chan and Kepert 2010). Dynamical models may be supplemented by statistical models to provide baseline guidance for evaluation or, in situations where dynamical models do not yet produce skillful forecasts, to provide actual operational guidance. For example, the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994), a simple linear regression scheme, is still one of the National Hurricane Center's best performing products for the prediction of TC intensity changes (DeMaria et al. 2014). Aberson (2014) developed a climatological “no-skill” baseline model that predicts the “phase” of the storm (i.e., whether it is in a tropical, subtropical, extratropical, dissipated, or “remnant low” stage of its life cycle). The model is a simple linear discriminant analysis scheme whose predictors match those of the National Hurricane Center's baseline models for track and intensity prediction. Our intent here is broadly similar to that of Aberson (2014), with the difference that the statistical model presented here is optimized for the prediction of the extratropical phase and ET.

Despite its important implications for risk assessment (Loridan et al. 2015), ET has received little attention in the development of TC hazard models for the (re)insurance industry (e.g., AIR Worldwide 2015) or in academia (e.g., Emanuel et al. 2006; Hall and Jewson 2007; Lee et al. 2018). The primary output of a hazard model is the annual probability of a storm intensity exceeding a given threshold at a specific location. Hazard models used in the industry (so-called catastrophe, or “cat” models) additionally incorporate vulnerability curves and insurance exposure data, which allow them to convert hazard intensity into an estimated level of financial loss. Due to the limited observed record, these models often use large ensembles of stochastically generated synthetic storms that resemble those in the historical data. In some, the storms' track and intensity evolution is guided by a few environmental parameters such as the state of the El Niño–Southern Oscillation, or seasonal or monthly average sea surface temperature. Integrating ET into hazard models thus requires a statistical prediction without real-time information about the environment.

Here, we introduce a statistical model that predicts how likely a TC is to be extratropical, given some information about the TC's present characteristics and environment. Two versions of the models are presented:

an “operational model” and a “hazard model.” The operational model makes forecasts at lead times up to four days, based on predictors that include real-time observations of the storm environment. This operational model is not intended to outperform dynamical models. Rather, its primary purpose is to serve as a “baseline model” that provides a benchmark for the evaluation of dynamical models. In other words, a dynamical model can be said to make skillful predictions if it beats the operational model presented here. There are numerous examples of forecast validations that define skill as an improvement over a benchmark given by a statistical model (e.g., Landsea and Knaff 2000; Vitart et al. 2010; Rashid et al. 2011; Eden et al. 2015). Note, however, that contrary to what its name may suggest, a baseline model can itself have actual skill when compared to a simple reference model such as climatology or persistence, or to a model that makes random predictions (and it does in this case, as will be shown in section 6). Sometimes, a so-called baseline model may even produce more accurate forecasts than a numerical weather prediction model (e.g., Cangialosi and Franklin 2016). The second version, the hazard model, is designed for use in TC risk assessments; it does not make true forecasts at all, but is diagnostic (i.e., it makes predictions at a lead time of 0 h) and works with a monthly mean representation of the environment.

This paper is organized as follows: section 2 introduces the datasets used to train and evaluate the regression models. Section 3 outlines the underlying mathematics and concepts of the models, and section 4 describes the model fitting and evaluation. The evaluation is based on the performance metrics presented in section 5. The results of the performance evaluation as well as some case studies are shown in section 6. Section 7 compares the models to cyclone phase space (CPS) diagnostics (Hart 2003), a well-established objective method to identify storms undergoing ET. Throughout the paper, we focus first on the operational model, and follow each segment with the corresponding analysis of the hazard model. The paper concludes with a summary of the findings (section 8).

2. Data: TC best track and reanalysis datasets

The features (predictors) are a combination of storm properties from 6-hourly TC best track data and environmental parameters derived from reanalysis data. All features are scaled to zero mean and unit variance. We use the best track archive of the National Hurricane Center (NHC) in the North Atlantic (NAT), and that of the Japan Meteorological Agency (JMA) in the western North Pacific (WNP). These datasets provide information on the position of the storm center, central pressure,

and the storm phase. The phase information is used to generate the predictand, which is a Boolean variable taking the value 1 if a storm is extratropical (EX) and 0 (non-EX) otherwise. Our data consist of 501 storms in the NAT and 966 storms in the WNP in the period 1979–2017.

Features taken from best track datasets are the same for the operational model and the hazard model, but the two models use different environmental features and reanalysis datasets.

a. Operational model

The 6-hourly environmental fields surrounding the TCs are taken from the Japanese 55-year Reanalysis (JRA-55; Kobayashi et al. 2015), which provides data on a regular latitude–longitude grid at a horizontal resolution of 1.25°. Geopotential height fields from JRA-55 are used to characterize the storms using the CPS (Hart 2003), which captures the physical structure of cyclones in terms of three parameters: the B parameter measures the asymmetry in the layer-mean temperature surrounding the cyclone, and two thermal wind (V_T) parameters assess whether the cyclone has a warm or cold core structure in the upper ($-V_T^U$) and lower ($-V_T^L$) troposphere (with the convention of the minus sign, positive values correspond to warm cores). A detailed explanation of the CPS can be found in Hart (2003).

Vertical wind shear is computed using area-weighted azimuthal mean Cartesian wind components at 200 and 850 hPa, following the method of Hanley et al. (2001). The azimuthal averaging, performed in five annular rings of 100-km width around the storm center, removes a symmetric vortex. Thus, the resulting areal-mean winds are measures of the environmental flow across the cyclone. The components of the areal-mean winds are given by

$$\langle u \rangle = \frac{1}{A} \sum_{r_i=r_1}^{r_5} \frac{\bar{u}_i + \bar{u}_{i-1}}{2} A_i,$$

$$\langle v \rangle = \frac{1}{A} \sum_{r_i=r_1}^{r_5} \frac{\bar{v}_i + \bar{v}_{i-1}}{2} A_i,$$

where u and v are the Cartesian wind components; A is the area of a circle of radius 500 km; A_i is the area of an annular ring with inner radius r_{i-1} and outer radius r_i ($r_0 = 0$ km, $r_1 = 100$ km, $\dots$, $r_5 = 500$ km); $\langle \cdot \rangle$ indicates an area average; and the overbar is an azimuthal average. The 200–850-hPa vertical shear (SHR) is then calculated as the magnitude of the vector difference between the area-averaged wind at the two pressure levels:

$$\text{SHR} = |\mathbf{u}_{200} - \mathbf{u}_{850}|, \quad (1)$$

$$\text{where } \mathbf{u}_{200} = \begin{bmatrix} \langle u_{200} \rangle \\ \langle v_{200} \rangle \end{bmatrix}, \quad \mathbf{u}_{850} = \begin{bmatrix} \langle u_{850} \rangle \\ \langle v_{850} \rangle \end{bmatrix}.$$

A further environmental feature is sea surface temperature (SST), which is averaged within a circle of radius 500 km around the storm center.

b. Hazard model

The hazard model is customized to function as a component in the Columbia Hazard model (CHAZ) developed by Lee et al. (2018) and therefore uses data from the European Centre for Medium-Range Weather Forecasts interim reanalysis (ERA-Interim; Dee et al. 2011) with a 0.75° horizontal grid spacing. For this application, monthly means of potential intensity (PI; Emanuel 1988) and 200–850-hPa vertical wind shear at the ERA-Interim grid point closest to the storm location are linearly interpolated from monthly means to daily values. PI is used instead of SST because CHAZ is designed for TC risk assessment in the present-day as well as in a future climate, and in the latter case, the relationship between SST and TC intensity is expected to change (Wing et al. 2007; Swanson 2008; Johnson and Xie 2010).

3. Logistic regression with elastic net

We apply a logistic regression model with elastic net regularization (Zou and Hastie 2005). This approach has the following advantages:

- Logistic regression produces probabilistic results and, as a generalized linear model, allows us to understand the impact of a feature on the predictand.
- The elastic net performs both regularization and feature selection, which enhances the prediction accuracy and interpretability of the model. It is often used when the number of features is large, perhaps even larger than the number of observations (e.g., Sokolov et al. 2016; Laing et al. 2018; Liu et al. 2018). In the application presented here, elastic net regularization was chosen out of preference for a simple model with as few predictors as possible, even though the initial number of features was not particularly large.
- The elastic net encourages a “grouping effect” (Zou and Hastie 2005), which is the property that highly correlated predictors tend to get equal coefficients (up to a change of sign if negatively correlated). This is an advantage over least absolute shrinkage and selection operator (LASSO) regularization (Tibshirani 1996), which also performs variable selection, but does not exhibit a grouping effect: it tends to pick only one of a

set of correlated features, such that the resulting selection of features may not include all features that provide predictive information.

At the core of logistic regression is the logistic function $h_{\theta}(\mathbf{x})$, which maps predicted values to probabilities:

$$h_{\theta}(\mathbf{x}) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_m \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}. \quad (2)$$

Given a feature vector $\mathbf{x}$ at time t , $h_{\theta}(\mathbf{x})$ is interpreted as $\hat{P}(\text{EX}|\mathbf{x})$ (i.e., as an estimate of the probability that the system will be extratropical at time $t + \Delta t$, where Δt is the lead time considered). If desired, this probability can be converted into a deterministic binary prediction by applying a decision threshold. In general, the optimal value of the decision threshold may be problem dependent. In this study, however, it was set to a fixed value of 0.5, as treating the threshold as a model parameter that is determined separately for each lead time only slightly improved model performance. Thus, the class is determined according to

$$\text{class} = \begin{cases} \text{EX}, & \text{if } h_{\theta}(\mathbf{x}) = \hat{P}(\text{EX}|\mathbf{x}) \geq 0.5 \\ \text{not EX}, & \text{if } h_{\theta}(\mathbf{x}) = \hat{P}(\text{EX}|\mathbf{x}) < 0.5 \end{cases}.$$

The vector θ contains the coefficients (weights), which are found by minimizing a cost function $J(\theta)$ defined as

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N \{\text{loss}[h_{\theta}(\mathbf{x}^{(i)}), y^{(i)}]\} + \lambda P_{\alpha}(\theta), \quad (3)$$

where N is the number of samples, $\mathbf{x}^{(i)}$ is the feature vector of the i th sample, $y^{(i)}$ is the i th predictand (1 for EX, 0 non-EX), and

$$\text{loss}[h_{\theta}(\mathbf{x}^{(i)}), y^{(i)}] = \begin{cases} -\log[h_{\theta}(\mathbf{x}^{(i)})], & y^{(i)} = 1 \\ -\log[1 - h_{\theta}(\mathbf{x}^{(i)})], & y^{(i)} = 0 \end{cases},$$

$$P_{\alpha}(\theta) = \sum_{j=1}^m [(1 - \alpha)\theta_j^2 + \alpha|\theta_j|], \quad \alpha \in [0, 1].$$

The second term on the right-hand side of (3), λP_{α} , is the elastic net term, which regularizes the model by penalizing a linear combination of the sum of the L2 norm and the L1 norm of each coefficient. The parameter α determines the relative contributions of the two norms. For $\alpha = 0$, the elastic net penalty is equivalent to the ridge penalty (Hoerl and Kennard 1970), and for $\alpha = 1$, it is equivalent to the LASSO penalty

(Tibshirani 1996). The parameter λ controls the strength of the regularization. The parameters α and λ are the so-called hyperparameters of the model. Their values are determined in a grid search over a range of possible combinations, each of which is evaluated in a 10-fold cross validation, as described in the next section.

4. Model fitting and evaluation

a. Operational model

The model is fit on a training set containing about 85% of the data, while the remaining 15% constitute the test set used to evaluate the model. Feature selection serves the purpose of producing a sparse model with better physical interpretability, and it is done here in the following way: On the training set, we perform cross-validated logistic regression with elastic net regularization for a range of different regularization parameters. As the regularization gets stronger, more and more features drop out (i.e., their coefficients shrink to zero). So-called “regularization paths,” plots of the feature coefficients as a function of the regularization strength, are shown in Fig. S1 in the online supplemental material. The higher the regularization strength required to drive a feature’s coefficient to zero, the more predictive skill that feature contributes (e.g., the most important feature in both ocean basins is sea surface temperature, which is the last one to drop out as the regularization strength increases). The combination of these regularization paths with the corresponding sequence of cross-validation performance scores (black dashed line in Fig. S1) describes a trade-off between model sparsity and performance. Determining the optimal set of features within this trade-off is partly a matter of preference for certain model properties (e.g., one may be willing to sacrifice some predictive skill for the benefit of having fewer predictors) and may depend on the particular application.

Based on the regularization paths in the NAT and the WNP, we determine a subset of 8 (out of 20) features that will be used for both basins. This feature selection is done once, for a lead time of 24 h. Model fits are then calculated using the training data for lead times from 0 to 96 h, in 6-hourly time steps, using only those eight features. Each model fit is the result of a grid search over a range of possible values for the hyperparameters α (the relative balance between the L1 and the L2 penalty) and λ (the overall regularization strength), where each combination of α and λ is evaluated in a 10-fold cross validation. In a 10-fold cross validation, the training data are divided into 10 folds. Out of the 10 folds, nine sets are used for training while the

remaining set is used for testing. The model is trained and tested 10 times, each time a new set is used as testing set while the remaining sets are used for training. The final result of the 10-fold cross-validation is then obtained by averaging the results of all 10 folds.

We apply this procedure separately for the WNP and the NAT data, which gives rise to two sets of feature weights for each lead time. To estimate confidence intervals for the feature weights, we apply a bootstrap procedure, in which we resample the training set 1000 times (with replacement), fit a model to each of these training sets, and collect the feature weights of each bootstrap replicate.

The evaluation of the model performance on the independent test set addresses three questions: how well does the model predict the future status (EX or not EX) of a storm, how well does it predict transitions (i.e., changes in status), and how well does it classify TCs into “ET storms” (storms that undergo ET at some point in their lifetimes) and “non-ET storms” (storms that do not undergo ET). Predicting the status is what the model learns in the training phase; evaluating the performance on this task in the test set is thus straightforward. The model’s ability to forecast transitions is evaluated by sliding a time window of 24 h along the predicted and the true time series of a storm’s “extratropical” status, computing the difference in the storm’s phase occurring between the last and the first time step of the time window. To account for the rare cases of tropical transitions (which would result in values of -1), the absolute value of the differences is taken. The sliding-window transformation results in time series that are 1 at time steps within 24 h of a status change, and 0 otherwise. For example, considering a lead time of 6 h, the transformed predictand is 1 if a phase transition occurs between 6 and 30 h, and 0 if no phase transition occurs within that time period.

We evaluate the model performance for 6-hourly lead times up to 96 h. To estimate the sensitivity of the performance to different random splits into training and test set, we resample the data to generate 50 bootstrap replicates for each lead time.

Finally, the classification into ET storms and non-ET storms is obtained from the model’s predictions at individual time steps. If, at any point in a TC’s lifetime, the model’s prediction switches from 0 (not extratropical) to 1 (extratropical), a TC is classified as an ET storm, and otherwise it is classified as a non-ET storm.

b. Hazard model

The hazard model is only fit for a lead time of 0 h, since forecasting a storm’s future phase is not necessary for the application in TC risk models. All features are

variables that are already used in CHAZ, and no feature selection is performed. Due to the model’s diagnostic nature, evaluating the prediction of status changes (in the way described in the previous section) is not possible as changes can only occur over nonzero time intervals.

Table 1 summarizes the features used for the operational model and the hazard model, and Table S1 lists all 20 features from which the features for the operational model were selected.

5. Performance metrics

The Matthews correlation coefficient (MCC; Matthews 1975) is regarded as a balanced skill metric for binary classification tasks, which can be used even if the two classes are of very different sizes. It is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \quad (4)$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. The MCC takes on values between -1 and 1 . A coefficient of 1 represents a perfect prediction, 0 is equivalent to a random prediction (e.g., as would be output by a classifier that generates predictions uniformly at random, or by a classifier that makes random predictions respecting the training set’s class distribution), and -1 indicates total disagreement between prediction and observation.

The Brier score loss (BSL; Brier 1950) is the mean squared difference between forecast probabilities and actual outcomes:

$$\text{BSL} = \frac{1}{N} \sum_{i=1}^N [h_{\theta}(\mathbf{x}^{(i)}) - y^{(i)}]^2, \quad (5)$$

Where N is the number of samples, $h_{\theta}(\mathbf{x}^{(i)})$ is the model’s prediction (a probability between 0 and 1) on sample $\mathbf{x}^{(i)}$, and $y^{(i)}$ is the true outcome (either 0 or 1). The BSL ranges between 0 and 1 , with smaller values indicating better predictions.

Two further metrics used in this study are precision and recall (e.g., Ting 2010), which are defined as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}. \end{aligned} \quad (6)$$

Precision is the ratio of correctly classified positive observations (here ET storms) to the total observations

TABLE 1. List of features used for the operational model and the hazard model.

Feature	Definition
Operational model	
B	CPS parameter B
H_a	Heading angle, measured clockwise from north
lat	Latitude of the storm center
P_o	Storm central pressure
SHR	Vertical wind shear (200–850 hPa) within a circle of radius 500 km around the storm center
SST	Mean sea surface temperature within a circle of radius 500 km around the storm center
T_s	Storm translational speed
$-V_T^U$	CPS parameter $-V_T^U$
Hazard model	
lat	Latitude of the storm center
Δlat	Change in lat during previous 12 h
P_o	Storm central pressure
ΔP_o	Change in P_o during previous 12 h
$\overline{\text{PI}}$	Potential intensity at storm location; daily value interpolated from monthly means
$\Delta\overline{\text{PI}}$	Change in $\overline{\text{PI}}$ during previous 12 h
SHR	Vertical wind shear (200–850 hPa) at storm location; daily value interpolated from monthly means
ΔSHR	Change in SHR during previous 12 h
T_s	Storm translational speed
ΔT_s	Change in T_s during previous 12 h

classified as positive. Applied to the classification into ET storms and non-ET storms explained in the previous section, it answers the question, “Of all storms the model declares to have undergone ET, what fraction actually did?” Recall is the ratio of correctly classified positive observations to all observations in the actual class—the corresponding question is “Of all true ET storms, what fraction does the model identify?”

The harmonic mean of precision and recall is called the F1 score (Chinchor 1992):

$$\text{F1} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Precision, recall, and F1 score reach their best value at 1 and worst value at 0.

6. Results

a. Operational model

1) FEATURE WEIGHTS

Figure 1a shows coefficients of the operational model for a lead time of 24 h. In both ocean basins, increasing

latitude and decreasing SST cause the largest increases in a cyclone’s odds of being extratropical at time $t + 24$ h. (Note that the coefficients of a logistic regression model describe the effect of a feature on the predictand in terms of odds ratios—a brief guide on how to interpret logistic regression coefficients is given in the appendix.) The influence of the SST is smaller in the NAT than in the WNP, possibly because TCs in the NAT experience a smaller spatial variability in SST as they tend to travel along mean SST gradients rather than across them (Foltz et al. 2018).

During ET, a storm loses its upper-level warm core and radial symmetry and becomes a cold-cored system with fronts, which is consistent with the signs of the $-V_T^U$ and B coefficients. Positive anomalies in translational speed and vertical wind shear, indicating a TC’s acceleration by midlatitude westerlies and interaction with a baroclinic environment, are associated with enhanced odds of being in the extratropical phase as well.

The heading angle H_a has a bimodal distribution in which negative anomalies correspond to directions ranging from north to northeastward. These directions are associated with TCs that recurve from their westward-moving tracks in the tropics into eastward ones in higher latitudes where they are more likely to undergo ET.

2) PERFORMANCE EVALUATION

Figure 2 shows the MCCs and BSLs for lead times from 0 to 96 h. For the status prediction (i.e., whether a storm will be extratropical or tropical), the model is most accurate for lead times up to one day, with maximum MCCs of about 0.9 in the WNP and 0.7 in the NAT (Fig. 2a). Note that the prediction at a lead time of 0 h is not trivial (and doesn’t necessarily have to achieve the highest score), since all predictions are made without knowledge of the storm’s current phase, and the feature selection has been made based on a lead time of 24 h. The scores in the two basins decrease at similar rates, dropping by about 0.3 between the 24- and 96-h forecasts. The performance difference between the two basins is confirmed by the lower (i.e., better) BSLs of the WNP model. It is noteworthy that the WNP model outperforms the NAT model even when they are trained on equal-sized training sets (recall that the two datasets differ in size by a factor of almost two due to the higher TC activity in the WNP).

The MCCs for the prediction of ET (computed on the time series resulting from applying a sliding time window of 24 h; Fig. 2c) are lower than those for the status prediction, which reflects the difficulty of predicting the precise timing of transitions. At a lead time of 24 h, the model predicts phase changes with an MCC of 0.4 in

FIG. 1. Model coefficients of (a) the operational model (for a lead time of 24 h) and (b) the hazard model in the North Atlantic (NAT) and the western North Pacific (WNP). The values (nondimensional) are shown on the x axis. The bars represent the median coefficients over 1000 bootstrapped replicates, and the error bars indicate the 95% confidence intervals. To allow for direct comparison of the coefficients between the NAT and the WNP despite the different underlying standard deviations, the WNP coefficients have been scaled by the ratio of the two standard deviations. The coefficients are defined in Table 1.

the NAT, and 0.6 in the WNP. Recall that a model that makes random predictions has an $MCC \approx 0$ (e.g., Boughorbel et al. 2017), and that an $MCC > 0$ thus means that a model has skill when compared to random predictions. In one-sample Wilcoxon signed rank tests (one test for each lead time), the sample mean MCC for the prediction of ET is significantly greater than zero for all lead times considered. However, even if the sample mean remains statistically distinguishable from zero for all lead times, Fig. 2c shows that the MCCs in both basins decrease rapidly from the 6-h lead time to the 48-h lead time forecast, and not much predictive skill is retained beyond a lead time of about three days.

The performance scores presented thus far are calculated on the basis of the model's predictions at individual time steps. These predictions can be used to assign each storm to one of two classes: ET storms,

which undergo ET at some point in their lifetimes, and non-ET storms, which do not undergo ET. The confusion matrices (2×2 contingency tables) in Fig. 3a summarize the performance of the resulting classification. In the NAT, the model classifies 36 of the test set storms as ET storms, 8 of which are false positives (precision = 0.78), and it correctly identifies 28 of the 35 true ET storms (recall = 0.80). In the WNP, 5 of the 63 storms classified as ET storms are false positives (precision = 0.92), and 58 of the 63 true ET storms are identified (recall = 0.92).

Distributions of the time difference between the predicted and actual occurrence of ET (Fig. 4a) are obtained from the true positive ET storms of the test set. For the calculation of these time differences, the forecast TC was considered to undergo ET at the first 24-h forecast with $\hat{P}(EX|\mathbf{x}) > 0.5$, regardless of whether the

FIG. 2. Performance of the operational model on the test set, for the NAT and WNP model: (left) Matthews correlation coefficient (MCC) and (right) Brier score loss (BSL). (a),(b) The performance scores for the prediction of the extratropical status, for lead times from 0 to 96 h. (c),(d) The scores for the prediction of status changes, for lead times from 6 to 96 h, using a sliding time window of 24 h. The markers (circles for the prediction of the extratropical status, triangles for the prediction of status changes) represent the medians of 50 bootstrapped replicates that were calculated for each lead time, with the shading representing the area between plus and minus one standard deviation.

forecast decreased in later forecasts. In both ocean basins, the median time difference is 6 h, and the difference between the predicted and the true ET time is less than 24 h in 93% of the cases in the NAT, and in 97% of the cases in the WNP.

3) EXAMPLES

Figure 5 shows the model's 24-h forecast for six storms in the test set (i.e., storms not used in the model development), whose meteorological histories we will outline in the following using information

from the TC reports provided by the NHC (available at <http://www.nhc.noaa.gov/data/tcr>) and the JMA (available at <http://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/annualreport.html>). The examples are chosen to illustrate different phase evolutions and degrees of model performance.

Karl (2004; Fig. 5a) was a Cape Verde hurricane that moved northwestward across the Atlantic before re-curving due to a baroclinic trough developing north of the storm. Continuing northeastward, Karl underwent ET and eventually made landfall in Norway as an

FIG. 3. Confusion matrices for (a) the 24-h forecast of the operational model and (b) the hazard model, in the NAT and WNP. The rows (y axis) correspond to the true classifications, and the columns (x axis) correspond to the model's classifications, such that the numbers of correct classifications (true positives and true negatives) are in the top-left and bottom-right quadrants, respectively, and the numbers of wrong classifications (false negatives and false positives) are in the top-right and bottom-left quadrants, respectively. The storms in the ET class undergo ET at some point during their lifetimes, while the storms in the non-ET class do not undergo ET. The matrices show the number of test set storms in each of the four possible classification outcomes, with the color indicating the respective percentage of all test set storms. Precision and recall are given below each matrix.

extratropical storm. Karl's predicted probability of being extratropical (yellow line) increases from nearly zero to one over the course of about 30 h, exceeding the threshold of 0.5 exactly 24 h before the NHC declares the system to be extratropical (blue line) (i.e., when plotted as a function of the verification date, the two lines cross the threshold of 0.5 at the same time). Thus, the model accurately forecasts Karl's phase evolution and transition time.

Similarly, the model captures the ET of Halong (2014; Fig. 5b), which took place over the Sea of Japan. Formed about 450 km east of Guam, Halong had rapidly intensified into a category 5 super typhoon, but weakened to tropical storm intensity by the time it made landfall as an extratropical system over the southern part of Japan.

FIG. 4. Difference between the predicted and the true ET time for all true positive ET storms of the test set, for (a) the 24-h forecast of the operational model and (b) the hazard model, in the NAT and WNP. The outer shape of the violin plot represents all possible results (except for time differences greater than three days, which are shown as crosses), with the vertical width indicating the frequency of occurrence. The thick black bar in the center represents the interquartile range, the thin black line extending from it represents the 95% confidence intervals, and the white dot is the median. Positive time differences indicate delayed model predictions.

Kit (1981; Fig. 5c) was a slow moving system that tracked to the northwest, passing just south of Guam, and then continued westward. Kit remained tropical throughout its lifetime, and the model indeed predicts the storm to remain tropical with almost 100% certainty.

Ivan (2004; Fig. 5d) is a rare example of a TC that undergoes ET and then undergoes "tropical transition" (e.g., Davis and Bosart 2004; Pezza and Simmonds 2008) back into a tropical system. Ivan first made landfall in Alabama, traveled through the eastern United States, looped back around the Atlantic Ocean and re-entered the Gulf of Mexico. Though not a perfect match, the model predicts the transition to the extratropical phase as well as the return to the tropical phase.

Helene (2006; Fig. 5e) formed from a tropical wave off the coast of Africa and gradually strengthened on its way northwest. Recurving along the northwest periphery of the subtropical ridge, Helene weakened and turned into a hybrid storm with both tropical and extratropical characteristics, featuring a deep warm core and an asymmetric, frontal-like appearance. However,

FIG. 5. The operational model's 24-h forecast for six examples from the test set: (a) Karl (2004, NAT) (b) Halong (2014, WNP), (c) Kit (1981, WNP), (d) Ivan (2004, NAT), (e) Helene (2006, NAT), and (f) Marilyn (1995, NAT). The yellow line shows the predicted probability of the cyclone being extratropical, and the blue line shows the true status (1: extratropical, 0: tropical) at time t .

the storm retained hurricane strength until becoming extratropical about 500 km northwest of the Azores. The model predicts Helene to become extratropical, but the transition is delayed and nonmonotonic. Consistent with the NHC's report, the $-V_T^U$ feature (not shown) indicates a persistent warm core in the upper troposphere that intensifies during the day before Helene

is declared extratropical, delaying ET in the model prediction.

After its formation in the eastern tropical Atlantic, Marilyn (1995; Fig. 5f) gradually intensified on its northwestward track, peaking as a category 3 hurricane shortly after hitting the Lesser Antilles at category 1 strength. Heading north past Bermuda, Marilyn

weakened and became extratropical. The model forecasts Marilyn to become increasingly extratropical over the three days before the NHC changes the status of the system, but does not predict a complete ET. Instead, the output probability drops back to nearly zero. The model's failure is the result of Marilyn abruptly turning south and then southeast after its ET, slowing down and meandering over an area with higher SST for the remainder of its lifetime.

The model's forecasts for these six storms are summarized in Fig. 6 as a function of lead time and verification date, in so-called chiclet diagrams (Carbin et al. 2016). The true phase evolution (as given by the best track datasets of the NHC and the JMA) is shown at the bottom of each panel. For any combination of lead time and verification date, the color shows the probability that the storm is extratropical at the verification date, as predicted by the model at the given lead time. For example, the chiclet diagram of Karl (2004) shows that the white color indicating the switch from tropical to extratropical is aligned with the actual transition up to a lead time of about 48 h. At longer lead times, the model still predicts an ET, but with increasing delay (as indicated by the positive slope of the line connecting the white blocks). Halong's (2014) ET is predictable on longer time scales: the white blocks remain vertically aligned up to a lead time of about three days, even though the transition becomes less "sharp." In contrast to Karl, the prediction of Halong's ET shifts to earlier times for longer forecast horizons.

Kit's (1981) tropical nature is highly predictable at all lead times considered here. Though present even in the 96-h forecast, the "double transition" of Ivan (2004) occurs with an increasing lag for longer lead times. For Helene (2006), the model's skill is almost independent of lead time, while Marilyn's (1995) phase evolution is, in fact, more accurately predicted at longer lead times, which show a longer extratropical phase (but still an incorrect reversal to a tropical status).

b. Hazard model

1) FEATURE WEIGHTS

In the hazard model, storm latitude and central pressure are the most important features for determining if a storm is in the extratropical phase (Fig. 1b). Changes over the previous 12 h contribute less to the prediction than do the current values of the features. Central pressure is given more weight in the hazard model than in the 24-h forecast of the operational model (Fig. 1a), especially in the NAT. The importance of central pressure depends strongly on the lead time considered: when the operational model is trained for the same diagnostic

(0-h lead time) task the hazard model is trained for, central pressure is the most important feature in the WNP, and also contributes predictive skill in the NAT (not shown). Over a 24-h time period, central pressure may vary too strongly to provide much predictive skill (e.g., in rapidly intensifying TCs, the central pressure falls by more than 40 hPa over the course of one day) (Holliday and Thompson 1979).

2) PERFORMANCE EVALUATION

The hazard model's performance scores for the prediction of the extratropical status in the NAT (MCC = 0.62, BSL = 0.05) and the WNP (MCC = 0.86, BSL = 0.02) are similar to those achieved by the operational model at short lead times.

The confusion matrices in Fig. 3b show that in the NAT, the model classifies 33 of the test set storms as ET storms, 5 of which are false positives (precision = 0.85), and it correctly identifies 28 of the 35 true ET storms (recall = 0.80). In the WNP, 9 of the 68 storms classified as ET storms are false positives (precision = 0.87), and 59 of the 63 true ET storms are identified (recall = 0.94). The model diagnoses 86% of all ET times in the NAT and 95% of those in the WNP with an error of less than 24 h (Fig. 4b).

Figure S2 shows the output of the hazard model for the six storms discussed for the operational model (Fig. 5). Similar to the operational model, the hazard model has difficulties diagnosing the phase evolutions of Ivan, Helene, and Marilyn, but captures the ETs of Karl and Halong as well as the purely tropical life cycle of Kit with high precision.

7. Comparison with CPS diagnostics

As mentioned in section 2, the CPS proposed by Hart (2003) has become a widely accepted diagnostic tool for defining ET, both in research and operational communities (Evans et al. 2017). However, the CPS is unable to distinguish between recurving TCs that undergo ET, and TCs that recurve without undergoing ET (Kofron et al. 2010). Further, the ET classification obtained from a CPS analysis depends on the dataset used to calculate the CPS parameters and is sensitive to choices of parameter thresholds (e.g., Hart 2003; Wood and Ritchie 2014; Bieli et al. 2019a,b). Evans et al. (2017) therefore recommend the continued evaluation of other ET classifiers. Note that an alternative model design could use CPS diagnostics (instead of the best track labels) to generate the predictand. Such a model was developed and tested, but, due to the reasons outlined above and the more reliable quality of the best track data in the NAT and the WNP since beginning of the satellite era

FIG. 6. Chiclet diagrams for six examples from the test set: (a) Karl (2004, NAT), (b) Halong (2014, WNP), (c) Kit (1981, WNP), (d) Ivan (2004, NAT), (e) Helene (2006, NAT), and (f) Marilyn (1995, NAT). The predicted probability of being extratropical (colored blocks) is shown as a function of the verification date (horizontal axis) and lead time (vertical axis). Below each chiclet diagram, a time series of the true (i.e., best track) phase evolution (0: not extratropical, 1: extratropical) is plotted with the y-axis inverted.

(e.g., [Holland and Webster 2007](#)), it was not included in this paper.

Here, we compare the performance of the logistic regression model introduced in this paper to that of the CPS analysis in [Bieli et al. \(2019b\)](#), hereafter [BCS19](#), who examine how well ET storms defined in the CPS agree with those defined in the best track records, using a global set of TCs from 1979 to 2017. Some of the statistics discussed here are not shown explicitly in [BCS19](#), but were reproduced here by the authors for the purpose of this comparison. In [BCS19](#), ET onset is defined as the first time a TC is either asymmetric ($B > 11$) or has a cold core ($-V_T^U < 0$ and $-V_T^L < 0$), and ET completion is when the second criterion is met. Two reanalysis datasets, JRA-55 and ERA-Interim, were used to calculate the CPS. To minimize false alarms due to thermal asymmetry in the early stage of a TC's lifetime, ET onset was only declared if a storm had a wind speed of at least 33 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$). The CPS analysis thus differs from the logistic regression model in that its definition of ET is not based on a statistical regression but on a combination of threshold criteria for the onset and completion of ET. Based on these criteria, [BCS19](#) classify each TC as an ET storm or as a non-ET storm, and the resulting CPS classification is evaluated against the best track extratropical labels using the MCC and the F1 score. [Table 2](#) compares their results to the test set performance scores obtained for the operational model in “diagnostic mode” (i.e., at a lead time of 0 h) and for the hazard model. In the NAT, both logistic regression models achieve higher F1 scores and MCCs than do the CPS diagnostics. In the WNP, the classifications from the logistic regression models and the JRA-55-based CPS diagnostics have almost equal scores, while those of the ERA-Interim-based CPS diagnostics are a bit lower.

[Figure 7](#) compares the timing errors made with respect to the true (i.e., best track) ET time: In [BCS19](#), the mean difference between the time of ET completion in the CPS and the true ET time for storms in the NAT is 10 h using JRA-55 and 32 h using ERA-Interim. The operational model and the hazard model have mean timing differences of 1 and 4 h, respectively. In the WNP, the CPS on average misses the true ET time by 5 h using JRA-55, and by 19 h using ERA-Interim, compared to time differences of 2 h for the operational model and 7 h for the hazard model. In both basins, the CPS timing errors have a higher variance. Thus, the logistic regression models diagnose the timing of ET with less bias and a lower variance than do the CPS diagnostics.

A separate analysis to evaluate the models' ability to distinguish recurring TCs that undergo ET from TCs that recur without undergoing ET was not performed.

TABLE 2. Performance scores for the classifications into ET storms and non-ET storms obtained from the operational model (at lead time 0 h), the hazard model, and from the CPS analysis in [Bieli et al. \(2019b\)](#), which was performed using JRA-55 as well as ERA-Interim data. The table shows the F1 score (F1) and the Matthews correlation coefficient (MCC) for the North Atlantic (NAT) and western North Pacific (WNP) basins.

	Operational model		Hazard model		CPS (JRA-55)		CPS (ERA-Interim)	
	F1	MCC	F1	MCC	F1	MCC	F1	MCC
NAT	0.85	0.73	0.88	0.79	0.77	0.57	0.76	0.55
WNP	0.90	0.80	0.91	0.82	0.90	0.81	0.86	0.73

When a storm moves into higher latitudes along a recurving track, the associated changes in heading angle and latitude increase its probability of being classified as extratropical. On the other hand, if the storm stays in a tropically favorable local environment characterized by high values of SST or PI, the model may still classify the storm as tropical. Thus, identifying recurving TCs that do not undergo ET is possible for the model in principle, but the performance would likely be lower than on the set of all (recurving and nonrecurving) storms.

8. Summary

We have introduced a logistic regression model to predict the extratropical transition (ET) of tropical cyclones (TCs), using predictors from best track and reanalysis datasets. Two versions of the model are presented: a baseline “operational model” that uses observed properties of the storm and its environment to forecast the storm's phase (extratropical or not extratropical), and a “hazard model” that diagnoses the phase without access to a representation of the storm's environment. The operational model was developed with a focus on predictive performance as well as physical interpretability and thus resides at the interface between machine learning and traditional statistics. It may be applied to provide baseline guidance for operational forecasts. The hazard model was developed with the purpose of integrating ET into statistical TC risk models used for hazard assessment.

Our findings can be summarized as follows:

- The operational model has skill in forecasting ET at lead times up to two days. At a lead time of 24 h, the Matthews correlation coefficient for the prediction of phase changes is 0.6 in the western North Pacific, and 0.4 in the North Atlantic.
- The model can be used to classify each TC of the test set as an ET storm (if it undergoes ET at some point in its lifetime) or as a non-ET storm (if it does not

FIG. 7. ET timing errors of the operational model (at lead time 0 h), the hazard model, and the CPS analysis in Bieli et al. (2019b), which was performed using JRA-55 as well as ERA-Interim data. Timing errors (shown on the x axis) are defined as the differences between the predicted ET times and the true (i.e., best track) ET times. Positive values thus indicate delayed predictions. The vertical lines denote the means of the distributions, the error bars represent values within one standard deviation about the mean, and the purple dots mark the location of the medians. Results are shown for the (top) NAT and (bottom) WNP.

undergo ET). Table 3 shows precision, recall, and the percentage of all true positive storms whose ET occurs within 24 h of the model's predicted ET time. Both the operational and the hazard model correctly identify 80% of true ET events in the North Atlantic and more than 90% of true ET events in the western North Pacific.

- For the 24-h lead time forecast, the most important predictors of the operational model are latitude and sea surface temperature. In the hazard model, the largest contributions come from latitude and storm central pressure.
- Using six storms from the test set, we illustrate that the model can predict the phase evolution of storms that undergo ET as well as of storms that remain tropical throughout their lifetimes, and that it also allows for tropical transitions (phase changes from extratropical to tropical).
- Used as an instantaneous diagnostic of a storm's tropical/extratropical status, the model performs about as well as the widely used cyclone phase space (CPS) in the western North Pacific and better than the CPS in the North Atlantic, and predicts the timings of the transitions better than CPS in both basins.

TABLE 3. Performance evaluation on the test set: precision and recall of the binary classification into ET storms and non-ET storms, as well as the percentage of all true positive storms whose ET occurs within 24 h of the model's predicted ET time. Values are shown for the 24-h forecast of the operational model (op), and for the hazard model (haz), in the North Atlantic (NAT) and the western North Pacific (WNP).

	Precision	Recall	ET timing error < 24 h
NAT, op, 24 h	0.78	0.80	93%
WNP, op, 24 h	0.92	0.92	97%
NAT, haz	0.85	0.80	86%
WNP, haz	0.87	0.94	95%

A next step will be to test the effectiveness of the operational model in real-time applications. In principle, the model is portable to any dataset that provides the environmental features used for model input. However, applying the model to a different dataset will require to refit the model, as the optimal coefficients likely depend on the dataset. Future work may also use the operational model as a benchmark for the validation of numerical weather prediction models' forecasts of ET. Regarding the model's integration into a TC risk assessment system, we hope that the ability to identify storms undergoing ET can provide a starting point for modeling the associated changes in wind or precipitation fields, which determine the risk of TC damage.

Acknowledgments. The first author gratefully acknowledges support from the NASA Modeling, Analysis, and Prediction (MAP) program, which provided the funding for this research under NASA Cooperative Agreement NNX15AJ05A. S.J.C. and A.H.S. acknowledge partial support from NOAA MAPP Grant NA16OAR4310079. The authors thank the three anonymous reviewers; their thoughtful comments helped improve the paper.

APPENDIX

Interpreting the Coefficients of a Logistic Regression

There is an important distinction between linear and logistic regression in the interpretation of the regression coefficients: In linear regression, a coefficient θ_i means that raising x_i by 1 causes an increase of θ_i in the expected value of the predictand y . In logistic regression, a coefficient θ_i means that raising x_i by 1 causes a multiplicative increase of e^{θ_i} in the odds that y occurs. In the following, we present a brief mathematical derivation of this result. A comprehensive treatment is given, for example, in Hosmer et al. (2013).

First, recall that in logistic regression, a given feature vector $\mathbf{x}$ is transformed by the logistic sigmoid function

$h_{\theta}(\mathbf{x})$ to give an estimate of the probability $\hat{P}$ that $\mathbf{x}$ is in class 1:

$$h_{\theta}(\mathbf{x}) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} = \hat{P}. \quad (\text{A1})$$

Letting $z = \theta^T \mathbf{x}$, $\hat{P}$ can be viewed as a function of z . It can easily be verified that the inverse of $\hat{P}$ is the logarithm of the odds, also known as the logit function:

$$\begin{aligned} \text{logit}(p) &= \log[\text{odds}(p)] = \log\left(\frac{p}{1-p}\right), \quad \text{and} \\ \text{logit}[\hat{P}(z)] &= \log\{\text{odds}[\hat{P}(z)]\} = z. \end{aligned} \quad (\text{A2})$$

The odds are a monotonic transformation of probabilities p and describe the chance of an event (here: “ $\mathbf{x}$ is in class 1”) as the ratio of the probability that the event does occur to the probability that it does not occur.

Exponentiating (A2) yields

$$e^z = e^{\theta^T \mathbf{x}} = \text{odds}[\hat{P}(z)]. \quad (\text{A3})$$

Thus, the coefficients θ determine the odds of class 1 for a given $\mathbf{x}$. To express the effect of an individual coefficient θ_i , suppose we increase feature x_i by one unit, holding all other features constant, and consider the odds ratio of x , $\text{OR}(x_i)$:

$$\text{OR}(x_i) = \frac{\text{odds}[\hat{P}(z)|z = \theta_0 + \theta_1 x_1 + \dots + \theta_i(x_i + 1) + \dots + \theta_n x_n]}{\text{odds}[\hat{P}(z)|z = \theta_0 + \theta_1 x_1 + \dots + \theta_i x_i + \dots + \theta_n x_n]}. \quad (\text{A4})$$

Using (A3), we can rewrite (A4) as

$$\text{OR}(x_i) = \frac{e^{\theta_0 + \theta_1 x_1 + \dots + \theta_i(x_i + 1) + \dots + \theta_n x_n}}{e^{\theta_0 + \theta_1 x_1 + \dots + \theta_i x_i + \dots + \theta_n x_n}} = e^{\theta_i}. \quad (\text{A5})$$

Thus, each exponentiated coefficient e^{θ_i} is the expected multiplicative change in the odds of class 1 for a unit increase in the corresponding predictor x_i , holding the other predictors constant. Note that this represents a *constant* effect that is independent of $\mathbf{x}$. While the odds resulting from a given $\mathbf{x}$ in (A3) can be converted to a probability, it is not possible to express the constant effect of logistic regression coefficients in terms of probabilities.

REFERENCES

- Aberson, S. D., 2014: A climatological baseline for assessing the skill of tropical cyclone phase forecasts. *Wea. Forecasting*, **29**, 122–129, <https://doi.org/10.1175/WAF-D-12-00130.1>.
- AIR Worldwide, 2015: The AIR hurricane model: AIR Atlantic tropical cyclone model v15.0.1 as implemented in Touchstone v3.0.0. AIR Worldwide Corporation Tech. Rep., 510 pp.
- Bieli, M., S. J. Camargo, A. H. Sobel, J. L. Evans, and T. Hall, 2019a: A global climatology of extratropical transition. Part I: Characteristics across basins. *J. Climate*, **32**, 3557–3582, <https://doi.org/10.1175/JCLI-D-17-0518.1>.
- , —, —, —, and —, 2019b: A global climatology of extratropical transition. Part II: Statistical performance of the cyclone phase space. *J. Climate*, **32**, 3583–3597, <https://doi.org/10.1175/JCLI-D-18-0052.1>.
- Boughorbel, S., F. Jarray, and M. El-Anbari, 2017: Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLOS ONE*, **12**, e0177678, <https://doi.org/10.1371/journal.pone.0177678>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, <https://doi.org/10.1175/MWR2905.1>.
- Cangialosi, J. P., and J. L. Franklin, 2016: 2015 National Hurricane Center Forecast Verification Report. National Hurricane Center Tech. Rep., 69 pp., http://www.nhc.noaa.gov/verification/pdfs/Verification_2015.pdf.
- Carbin, G. W., M. K. Tippett, S. P. Lillo, and H. E. Brooks, 2016: Visualizing long-range severe thunderstorm environment guidance from CFSv2. *Bull. Amer. Meteor. Soc.*, **97**, 1021–1031, <https://doi.org/10.1175/BAMS-D-14-00136.1>.
- Chan, J. C. L., and J. D. Kepert, 2010: *Global Perspectives on Tropical Cyclones: From Science to Mitigation*. World Scientific, 448 pp.
- Chinchor, N., 1992: MUC-4 evaluation metrics. *Proc. Fourth Conf. on Message Understanding*, McLean, VA, Association for Computational Linguistics, 22–29.
- Davis, C. A., and L. F. Bosart, 2004: The TT problem: Forecasting the tropical transition of cyclones. *Bull. Amer. Meteor. Soc.*, **85**, 1657–1662, <https://doi.org/10.1175/BAMS-85-11-1657>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/QJ.828>.
- DeMaria, M., and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- , C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, <https://doi.org/10.1175/BAMS-D-12-00240.1>.
- Eden, J., G. J. van Oldenborgh, E. Hawkins, and E. B. Suckling, 2015: A global empirical system for probabilistic seasonal climate prediction. *Geosci. Model Dev.*, **8**, 3947–3973, <https://doi.org/10.5194/gmd-8-3947-2015>.
- Emanuel, K. A., 1988: The maximum intensity of hurricanes. *J. Atmos. Sci.*, **45**, 1143–1155, [https://doi.org/10.1175/1520-0469\(1988\)045<1143:TMIOH>2.0.CO;2](https://doi.org/10.1175/1520-0469(1988)045<1143:TMIOH>2.0.CO;2).

- , S. Ravela, E. Vivant, and C. Risi, 2006: A statistical deterministic approach to hurricane risk assessment. *Bull. Amer. Meteor. Soc.*, **87**, 299–314, <https://doi.org/10.1175/BAMS-87-3-299>.
- Evans, C., and Coauthors, 2017: The extratropical transition of tropical cyclones. Part I: Cyclone evolution and direct impacts. *Mon. Wea. Rev.*, **145**, 4317–4344, <https://doi.org/10.1175/MWR-D-17-0027.1>.
- Fogarty, C., 2010: Forecasting extratropical transition. *Proc. Seventh Int. Workshop on Tropical Cyclones*, St. Gilles Les Bains, La Reunion, France, World Meteorological Organization, 2.5, http://www.wmo.int/pages/prog/arep/wwrp/tmr/otherfileformats/documents/2_5.pdf.
- Foltz, G. R., K. Balaguru, and S. Hagos, 2018: Interbasin differences in the relationship between SST and tropical cyclone intensification. *Mon. Wea. Rev.*, **146**, 853–870, <https://doi.org/10.1175/MWR-D-17-0155.1>.
- Hall, T. M., and S. Jewson, 2007: Statistical modelling of North Atlantic tropical cyclone tracks. *Tellus*, **59A**, 486–498, <https://doi.org/10.1111/j.1600-0870.2007.00240.x>.
- Hanley, D., J. Molinari, and D. Keyser, 2001: A composite study of the interactions between tropical cyclones and upper-tropospheric troughs. *Mon. Wea. Rev.*, **129**, 2570–2584, [https://doi.org/10.1175/1520-0493\(2001\)129<2570:ACSOTI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2570:ACSOTI>2.0.CO;2).
- Hart, R. E., 2003: A cyclone phase space derived from thermal wind and thermal asymmetry. *Mon. Wea. Rev.*, **131**, 585–616, [https://doi.org/10.1175/1520-0493\(2003\)131<0585:ACPSDF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<0585:ACPSDF>2.0.CO;2).
- , and J. L. Evans, 2001: A climatology of the extratropical transition of Atlantic tropical cyclones. *J. Climate*, **14**, 546–564, [https://doi.org/10.1175/1520-0442\(2001\)014<0546:ACOTET>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<0546:ACOTET>2.0.CO;2).
- Hoerl, A. E., and R. W. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67, <https://doi.org/10.1080/00401706.1970.10488634>.
- Holland, G. J., and P. J. Webster, 2007: Heightened tropical cyclone activity in the North Atlantic: Natural variability or climate trend? *Philos. Trans. Roy. Soc. London*, **365A**, 2695–2716, <https://doi.org/10.1098/rsta.2007.2083>.
- Holliday, C. R., and A. H. Thompson, 1979: Climatological characteristics of rapidly intensifying typhoons. *Mon. Wea. Rev.*, **107**, 1022–1034, [https://doi.org/10.1175/1520-0493\(1979\)107<1022:CCORIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1979)107<1022:CCORIT>2.0.CO;2).
- Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant, 2013: *Applied Logistic Regression*. 2nd ed. John Wiley & Sons, 528 pp.
- Johnson, N. C., and S.-P. Xie, 2010: Changes in the sea surface temperature threshold for tropical convection. *Nat. Geosci.*, **3**, 842–845, <https://doi.org/10.1038/ngeo1008>.
- Jones, S. C., and Coauthors, 2003: The extratropical transition of tropical cyclones: Forecast challenges, current understanding, and future directions. *Wea. Forecasting*, **18**, 1052–1092, [https://doi.org/10.1175/1520-0434\(2003\)018<1052:TETOTC>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<1052:TETOTC>2.0.CO;2).
- Keller, J. H., S. Jones, J. Evans, and P. Harr, 2011: Characteristics of the TIGGE multimodel ensemble prediction system in representing forecast variability associated with extratropical transition. *Geophys. Res. Lett.*, **38**, L12802, <https://doi.org/10.1029/2011GL047275>.
- , and Coauthors, 2019: The extratropical transition of tropical cyclones. Part II: Interaction with the midlatitude flow, downstream impacts, and implications for predictability. *Mon. Wea. Rev.*, **147**, <https://doi.org/10.1175/MWR-D-17-0329.1>.
- Kitabatake, N., 2011: Climatology of extratropical transition of tropical cyclones in the western North Pacific defined by using cyclone phase space. *J. Meteor. Soc. Japan*, **89**, 309–325, <https://doi.org/10.2151/JMSJ.2011-402>.
- Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, <https://doi.org/10.2151/JMSJ.2015-001>.
- Kofron, D. E., E. A. Ritchie, and J. S. Tyo, 2010: Determination of a consistent time for the extratropical transition of tropical cyclones. Part I: Examination of existing methods for finding “ET time.” *Mon. Wea. Rev.*, **138**, 4328–4343, <https://doi.org/10.1175/2010MWR3180.1>.
- Laing, E. E., C. S. Möller-Levet, D.-J. Dijk, and S. N. Archer, 2018: Identifying and validating blood mRNA biomarkers for acute and chronic insufficient sleep in humans: A machine learning approach. *Sleep*, **42**, zsy186, <https://doi.org/10.1093/SLEEP/ZSY186>.
- Landsea, C. W., and J. A. Knaff, 2000: How much skill was there in forecasting the very strong 1997–98 El Niño? *Bull. Amer. Meteor. Soc.*, **81**, 2107–2119, [https://doi.org/10.1175/1520-0477\(2000\)081<2107:HMSWTI>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<2107:HMSWTI>2.3.CO;2).
- Lee, C.-Y., M. K. Tippett, A. H. Sobel, and S. J. Camargo, 2018: An environmentally forced tropical cyclone hazard model. *J. Adv. Model. Earth Syst.*, **10**, 223–241, <https://doi.org/10.1002/2017MS001186>.
- Liu, J., G. Liang, K. D. Siegmund, and J. P. Lewinger, 2018: Data integration by multi-tuning parameter elastic net regression. *BMC Bioinf.*, **19**, 369, <https://doi.org/10.1186/s12859-018-2401-1>.
- Loridan, T., S. Khare, E. Scherer, M. Dixon, and E. Bellone, 2015: Parametric modeling of transitioning cyclone wind fields for risk assessment studies in the western North Pacific. *J. Appl. Meteor. Climatol.*, **54**, 624–642, <https://doi.org/10.1175/JAMC-D-14-0095.1>.
- Matthews, B. W., 1975: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451, [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9).
- Pezza, A. B., and I. Simmonds, 2008: Large-scale factors in tropical and extratropical cyclone transition and extreme weather events. *Ann. N.Y. Acad. Sci.*, **1146**, 189–211, <https://doi.org/10.1196/annals.1446.005>.
- Rashid, H. A., H. H. Hendon, M. C. Wheeler, and O. Alves, 2011: Prediction of the Madden-Julian Oscillation with the POAMA dynamical prediction system. *Climate Dyn.*, **36**, 649–661, <https://doi.org/10.1007/s00382-010-0754-x>.
- Sokolov, A., D. E. Carlin, E. O. Paull, R. Baertsch, and J. M. Stuart, 2016: Pathway-based genomics prediction using generalized elastic net. *PLOS Comput. Biol.*, **12**, e1004790, <https://doi.org/10.1371/journal.pcbi.1004790>.
- Swanson, K. L., 2008: Nonlocality of Atlantic tropical cyclone intensities. *Geochem. Geophys.*, **9**, <https://doi.org/10.1029/2007GC001844>.
- Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc.*, **58B**, 267–288.
- Ting, K. M., 2010: Precision and recall. *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Springer, 781 pp.
- Vitart, F., A. Leroy, and M. C. Wheeler, 2010: A comparison of dynamical and statistical predictions of weekly tropical cyclone activity in the Southern Hemisphere. *Mon. Wea. Rev.*, **138**, 3671–3682, <https://doi.org/10.1175/2010MWR3343.1>.
- Wing, A. A., A. H. Sobel, and S. J. Camargo, 2007: Relationship between the potential and actual intensities of tropical cyclones on interannual time scales. *Geophys. Res. Lett.*, **34**, L08810, <https://doi.org/10.1029/2006GL028581>.
- Wood, K. M., and E. A. Ritchie, 2014: A 40-year climatology of extratropical transition in the eastern North Pacific. *J. Climate*, **27**, 5999–6015, <https://doi.org/10.1175/JCLI-D-13-00645.1>.
- Zou, H., and T. Hastie, 2005: Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc.*, **67B**, 301–320, <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.