A Statistical Investigation of the Dependence of Tropical Cyclone Intensity Change on the Surrounding Environment

NING LIN AND RENZHI JING
Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

YUYAN WANG
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey

EMMI YONEKURA
Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

JIANGQING FAN
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey

LINGZHOU XUE
Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 25 September 2016, in final form 14 March 2017)

ABSTRACT

A progression of advanced statistical methods is applied to investigate the dependence of the 6-h tropical cyclone (TC) intensity change on various environmental variables, including the recently developed ventilation index (VI). The North Atlantic (NA) and western North Pacific (WNP) observations from 1979 to 2014 are used. As a first step, a model of the intensity change is developed as a linear function of 13 variables used in operational models, obtaining statistical $R^2$ values of 0.26 for NA and 0.3 for WNP. Statistical variable selection techniques are then applied to significantly reduce the number of predictors (to 5–11), while keeping similar $R^2$ values with linear or nonlinear models. Further reduction of the number of predictors (to 5–7) and significant improvement of $R^2$ (0.41–0.53) are obtained with mixture modeling, indicating that the dependence of TC intensification on the environment is nonhomogeneous. Applying VI as the environmental predictor in the mixture modeling gives $R^2$ results (0.41–0.74) similar to or better than those with more environmental variables, confirming that VI is a dominant environmental variable, although its effect on TC intensification is quite heterogeneous. However, the overall predictive $R^2$ results of the mixture models are relatively low (<0.3), as the considered environmental variables have limited predictability for the occurrence of extreme/rapid intensification. Finally, nonparametric regression with six predictors [current intensity, previous intensity change, the three components of VI (maximum potential intensity, shear, and entropy deficit), and 200-hPa zonal wind] performs relatively well with predictive $R^2$ values of 0.37 for NA and 0.36 for WNP. The predictability of these statistical models may be further improved by adding oceanic and inner-core process predictors.

1. Introduction

The intensity of a tropical cyclone (TC) has been one of the most difficult aspects of TCs to model, as TC intensification is a complex dynamic process depending on
many components of the environment. Based on empirical relationships, the Statistical Hurricane Intensity Prediction Scheme (SHIPS) and Statistical Typhoon Intensity Prediction Scheme (STIPS) models have been built for operational weather forecasts of TC intensity (Knaff et al. 2005; DeMaria and Kaplan 1999; DeMaria et al. 2005). The models show that intensification depends on current and previous TC characteristics (e.g., translational speed, current intensity) and various environmental factors [e.g., maximum potential intensity (MPI), vertical wind shear, upper-level winds, and ocean heat content], though the empirical relationships change in different ocean basins. A simplified operational forecast scheme has also been developed to relate TC intensification to a reduced number of environmental variables including MPI, wind shear, and convective instability (DeMaria 2009). Emanuel et al. (2004) use a simple dynamic model to show that TC intensity depends significantly on MPI and other environmental factors, including wind shear, upper-ocean thermal structure, bathymetry in shallow regions, and land moisture and temperature. The study further notes that TCs also have internal variability of intensity through eyewall replacement cycles, which can be independent of the environment. Moreover, during their life cycles, many TCs undergo rapid intensification (RI; Kaplan and DeMaria 2003), which has been shown to significantly affect the climatology of the maximum lifetime intensity of TCs (Lee et al. 2016). Previous studies have attributed RI to inner-core, oceanic, and large-scale environmental processes (e.g., Kaplan and DeMaria 2003; Kaplan et al. 2010, 2015). However, the physical mechanism of RI and its dependence on the environment are not completely understood, and the operational prediction of RI has been a great challenge (Elsberry et al. 2007).

TC intensity plays a large role in the damage potential of a storm (e.g., Kantha 2006, 2008; Chavas et al. 2013). TC intensity modeling is not only critical in real-time operational forecasting but also a key component of long-term risk assessment. Within a risk assessment framework, a large number of synthetic TCs, generated with Monte Carlo (MC) methods, are often needed to evaluate the risk posed to a specific region (e.g., Lin and Emanuel 2006). The intensity is simulated along the generated storm track (e.g., Vickery et al. 2000; Emanuel et al. 2006; Hall and Jewson 2007; Yonekura and Hall 2011), possibly using the estimated environmental fields along the track; the environmental fields are estimated from observed or climate-model-projected datasets. Some TC intensity models that are used operationally may be applied to risk assessment. For such purposes, it is best to have both a computationally simple model and, more importantly, one that requires fewer environmental variable predictors, as some of the environmental variables may not be available in climate-model-projected datasets. The Coupled Hurricane Intensity Prediction System (CHIPS; Emanuel et al. 2004) is an example of a model used for operational forecasting that can also be applied to risk assessment (Emanuel et al. 2006). CHIPS is a relatively simple deterministic model based on TC physics. In contrast, a statistical model based on empirical relationships from observations of TCs and their environment may also be applied. Ideally, such a model relies on a few environmental predictors and captures nearly the full range of intensity changes.

The dependence on the environment, however, is critical, as it is key to a TC intensity model that will be used for long-term projection and risk analysis within the context of climate variation and climate change. Most previous work on statistical modeling of TC intensity for risk analysis relies very little on the environment but heavily on storm characteristics. The simplest statistical models for intensity use only current and previous TC intensities (i.e., persistence) as predictors. One example is the James and Mason (2005) model for TC central pressure, which is modeled as an autoregressive process. (Persistence is a very strong predictor, especially for short time step predictions.) The Vickery et al. (2000) model takes the persistence model a step further by using current and previous sea surface temperatures (SSTs) as predictors for statistical simulations of relative intensity—the TC central pressure normalized by the minimum central pressure that mean seasonal climatic conditions would allow. In an updated model, Vickery et al. (2009) add dependence on vertical wind shear, and the relative intensity is determined using the tropopause temperature and SST adjusted by a 1D ocean model (Emanuel et al. 2006). Recently, Lee et al. (2015) have developed a TC intensity model that reduces the SHIPS multiple linear regression model to its key environmental predictors: MPI minus current intensity, vertical wind shear, atmospheric stability, and 200-hPa divergence.

The main objective of this work is to further explore the dependence of TC intensification (including deintensification) on the environment and to develop new statistical models of TC intensity. We first build regression models incorporating environmental predictors that were previously considered to study their influence on TC intensification. However, as demonstrated by Lee et al. (2015), the previously considered variables are limited in fully capturing the climatology of TC intensity. Thus, we also explore a new variable from a recent advancement in our understanding of TC physics. Tang and Emanuel (2012) present an environmental predictor for intensification, the ventilation index (VI). They show that the ventilation of low-entropy air into a TC via vertical wind shear plays an important role in TC
formation and intensification. VI is a nonlinear combination of critical environmental variables: it is the product of vertical wind shear and the entropy deficit between the environment and the inner storm core, divided by MPI. The entropy deficit characterizes the mid-level horizontal entropy difference between the storm and environment relative to the vertical entropy difference between the sea surface and the air above it. Tang and Emanuel (2012) derive a relationship between intensity change normalized by MPI as a joint function of VI and intensity normalized by MPI, showing that positive intensification and negative intensification occur in distinct phases of the VI-normalized intensity space; VI is a potentially valuable predictor for modeling TC intensity change. As such, we construct TC intensification regression models using VI as a sole environmental predictor. As VI is a nondimensional quantity with strong theoretical relationships to nondimensional intensity and intensification (Tang and Emanuel 2012), we also construct fully nondimensional VI models. The models based on VI, as well as on its component variables, are compared with the models based on previously considered environmental variables.

Recent advancement in statistical analysis of complex, large datasets also motivates us to explore new regression methods applied to TC intensity modeling. First, similar to Lee et al. (2015), we reduce the SHIPS and STIPS models by statistically identifying the most important environmental predictors and developing their associated linear models. However, instead of using the forward selection procedure of Lee et al. (2015) and others (DeMaria and Kaplan 1999; Knaff et al. 2005), which adds individual predictors in a stepwise manner, we add a constraint to the coefficient estimation process to select important predictors simultaneously from a large pool of potential predictors. Second, we construct generalized additive models, which allow nonlinear dependence of the TC intensification on the storm and environmental variables. The previous studies (DeMaria and Kaplan 1999; Knaff et al. 2005) have also used some nonlinear predictors (e.g., MPF), indicating that there may exist nonlinear relationships with predictors. Here, we attempt to test the nonlinearity in all selected variables in a more systematic way. Third, Musgrave et al. (2012) note that the life cycle intensity of a TC may be divided into distinct regimes, meaning not all intensification observations should be treated the same. The idea of multiple regimes motivates us to treat TC intensification as a heterogeneous mixture in a finite mixture regression. Further, in addition to the parametric models with linear or specified nonlinear and heterogeneous structures, we also apply nonparametric modeling to explore the predictive potential of the variables of interest. These different statistical analysis approaches will greatly extend previous linear regression methods in order to improve TC intensity forecasting and risk analysis.

The four main phases of our statistical analysis are: variable selection, linear and nonlinear regression, finite mixture regression, and nonparametric analysis of selected variables. These statistical analyses are applied to a large number of previously considered environmental variables as well as to VI as the sole environmental predictor for TC intensification. Section 2 describes the data and the different statistical methods that are applied in the analysis. We highlight the progressions from a simple linear model to more complex approaches and discuss their advantages. Section 3 presents and discusses results from each phase of the analysis, performing a comparison of statistical performance, environmental predictor reductions, and interpretability. Section 4 discusses prediction perspectives, especially for the finite mixture modeling and nonparametric modeling. In section 5, we summarize our key findings and new insights regarding the statistical modeling of TC intensity change.

2. Methodology

a. Data and preprocessing

Most of the potential predictors, including TC and environmental variables, are similar to those used for both the SHIPS and STIPS models (DeMaria and Kaplan 1999; Knaff et al. 2005). A complete list of the notations, calculations, and data sources of the considered variables appear in Table 1. TC data are taken from the International Best Track Archive for Climate Stewardship (IBTrACS) WMO archive (Knapp et al. 2010). The TC data include 6-hourly latitude and longitude positions and 10-min maximum sustained wind speeds at 10 m above the surface. The location data are used to calculate the translational speed (TSPD) of the storm. Estimated surface background wind (Lin and Chavas 2012) is subtracted from the observed maximum wind to obtain the TC maximum wind (V). The 6-h change in the TC wind intensity from current to next time step, denoted by DV, is the intensification we model in this study. The 6-h step change in wind intensity from the previous time step is represented as DV\(_p\). We note that the SHIPS and STIPS models, as well as the Lee et al. (2015) model, consider various forecast windows from 12 to 120 h, as required for real-time forecasting. These analyses show that the statistical performance of the models decreases for shorter forecast windows, likely because of an increasing ratio of the data error to intensity change as the time interval becomes shorter. Nevertheless, here we focus on 6-h intensity changes in line with the statistical approach for risk analysis, which simulates both the track and intensity at 6-hourly intervals to mimic the format of the best track
observations (Vickery et al. 2000, 2009; Hall and Yonekura 2013). The methods developed here can be similarly applied to longer-time-interval analyses for further investigation and for real-time forecasting applications.

Environmental data are taken from the European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim) with a resolution of $0.75^\circ \times 0.75^\circ$ (Dee et al. 2011). The environmental variables include SHR, the vertical wind shear or difference between the 850- and 200-hPa level winds; USHR, the zonal component of the vertical wind shear; LSHR, the latitude-considered wind shear; T200, U200, and D200, the 200-hPa temperature, zonal wind, and divergence, respectively; RHHI, the relative humidity in the layer between 300 and 500 hPa; and Z850, the 850-hPa absolute vorticity. More complex combinations of environmental variables include MPI and VI. The MPI is calculated following Emanuel (1995) and Bister and Emanuel (1998), and VI [including its component variable entropy deficit (ED)] is calculated according to Tang and Emanuel (2012). In some analyses involving VI, the TC intensity is normalized by the maximum potential intensity ($V_{\text{MPI}}$), with 6-h intensification to the next step denoted as $d(V_{\text{MPI}})$ and from the previous step denoted as $d(V_{\text{MPI}})_p$. To account for the tendency toward spatial and temporal changes, we use the next-step values of the environmental variables to predict the change in intensity from the current to the next step. The next-step environmental values are available for risk analysis as the environment and storm track are first simulated. The current step environmental variables can be used as predictors, if the next-step environment is not first predicted in real-time forecasting applications.

Sensitivity analyses show that the results are similar when using current or next-step environmental values in our short-time-interval prediction (6 h). In this study we do not incorporate oceanic variables to directly account for the ocean’s negative impact on storm intensification. Since 2004, SHIPS has started to incorporate oceanic heat content (OHC) estimates inferred from satellite altimetry observations; however, the oceanic data are only available for a limited time period and over a limited part of the Atlantic basin.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Name</th>
<th>Source/calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V$</td>
<td>Max sustained wind speed</td>
<td>IBTRACS WMO</td>
</tr>
<tr>
<td>$DV_p$</td>
<td>6-h change in $V$ from previous time step</td>
<td>$DV_p = V_t - V_{t-1}$ ($t$ is the current time)</td>
</tr>
<tr>
<td>JDAY</td>
<td>Julian day from peak activity day</td>
<td>IBTRACS WMO day of days away from the 253rd day in NA and 248th day in WNP</td>
</tr>
<tr>
<td>TSPD</td>
<td>Translational speed</td>
<td>IBTRACS WMO positions</td>
</tr>
<tr>
<td>SHR</td>
<td>Vertical wind shear</td>
<td>ERA-Interim; difference of daily 850- and 200-hPa wind vectors averaged over 200–800-km annulus around the storm center</td>
</tr>
<tr>
<td>USHR</td>
<td>Zonal component of vertical wind shear</td>
<td>Zonal component of SHR</td>
</tr>
<tr>
<td>LSHR</td>
<td>Lat-considered shear</td>
<td>IBTRACS WMO for storm lat</td>
</tr>
<tr>
<td>T200</td>
<td>200-hPa atmospheric temperature</td>
<td>ERA-Interim; daily value averaged over 200–800-km annulus around the storm center</td>
</tr>
<tr>
<td>U200</td>
<td>200-hPa zonal wind</td>
<td>ERA-Interim; daily value averaged over 200–800-km annulus around the storm center</td>
</tr>
<tr>
<td>D200</td>
<td>200-hPa divergence</td>
<td>ERA-Interim; daily value averaged over 1000-km-radius area around the storm center</td>
</tr>
<tr>
<td>RHHI</td>
<td>High-level relative humidity</td>
<td>ERA-Interim; daily value averaged from 300 to 500 hPa and within 200–800-km annulus around the storm center</td>
</tr>
<tr>
<td>Z850</td>
<td>850-hPa absolute vorticity</td>
<td>ERA-Interim; daily value averaged over 1000-km-radius area around the storm center</td>
</tr>
<tr>
<td>MPI</td>
<td>Max potential intensity in max sustained wind</td>
<td>Calculation following Emanuel (1995) and Bister and Emanuel (1998)</td>
</tr>
<tr>
<td>VI</td>
<td>Ventilation index</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>ED</td>
<td>Entropy deficit</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>$V_{\text{MPI}}$</td>
<td>Normalized wind intensity</td>
<td>$V_{\text{MPI}}$</td>
</tr>
<tr>
<td>$d(V_{\text{MPI}})_p$</td>
<td>6-h change in $V_{\text{MPI}}$ from previous time step</td>
<td>$d(V_{\text{MPI}})<em>p = (V</em>{\text{MPI}})<em>t - (V</em>{\text{MPI}})_{t-1}$ ($t$ is the current time)</td>
</tr>
</tbody>
</table>
(DeMaria et al. 2005). Lee et al. (2015) consider the ocean temperature averaged over the top 100 m (Price 2009) but find that it is not a significant predictor. How to better account for the negative impact of the ocean in statistical TC intensity modeling warrants further research.

We use data from the years 1979–2014 for both the North Atlantic (NA) and western North Pacific (WNP). The data years are chosen to capture observations from the era of satellite technology as well as to include possible decadal oscillations. We start by developing SHIPS- and STIPS-like models for comparison. Differences from the original SHIPS and STIPS published models are expected as a result of the increased number of data years in our analysis: the DeMaria and Kaplan (1999) SHIPS model was built using 1989–96 observations, and the Knaff et al. (2005) STIPS model was built using 1997–2002 observations. Here, we use the environmental data from the ERA-Interim, whereas the SHIPS and STIPS models were constructed using operational synoptic datasets (DeMaria and Kaplan 1999; Knaff et al. 2005). We adhere to the spatial averaging of environmental variables in SHIPS that attempts to exclude the influence of the TC on the observed environmental variables.

The domains of analysis are 0°–30°N and 100°–12°W for NA and 0°–30°N and 95°E–180° for WNP. Only the 6-h observations that are over the ocean are kept for the analysis because TC intensity is known to behave differently over land. The observations that have missing values for any variable are removed. For example, if an observation of $DV$ is missing the value for U200, the observation is not used in the regression. The filtered datasets include 7171 observations for NA and 13,180 observations for WNP. In the statistical analysis, variables are first standardized (by subtracting the mean of all observations and dividing by the standard deviation), except for the nondimensional VI models.

b. Statistical methods

The models predict the intensification $DV$ (predictand) given the environmental and storm variables (predictors). Initially, 13 TC and environmental variables informed by the SHIPS and STIPS models [listed in Table 1, except VI, ED, $V/\text{MPI}$, and $d(V/\text{MPI})_p$,] are used in a multiple linear regression model to serve as a baseline for comparison. The first stage of more complex models uses variable selection to reduce the model to only the most important predictors. Generalized additive modeling is applied next to allow for nonlinear relationships between the selected variables and the predictand. Then, the observations are considered to be a heterogeneous mixture and are classified into groups so that each has its own selected predictors and statistical properties. This progression of analyses is also used to test VI as the sole environmental predictor, with only storm intensity variables (i.e., $V$ and $DV_p$) as the other possible predictors. In addition, nondimensional analyses with VI are performed. Finally, the nonparametric method is applied to the analysis of VI, its component variables (MPI, SHR, ED), and other variables highlighted by the variable selection analysis. The following five subsections describe the statistical methods applied in these analyses.

1) ORDINARY LEAST SQUARES (OLS)

In the linear regression with OLS, the squared residual error is minimized to estimate the model parameters $\beta_j$ and $\alpha$:

$$Y = \alpha + \sum_{j=1}^{m} \beta_j X_j + \epsilon, \; \epsilon \sim N(0, \sigma^2) \quad \text{and} \quad (1a)$$

$$\begin{align*}
\hat{(\alpha, \beta)} = \argmin_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \alpha - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2.
\end{align*} \quad (1b)$$

The predictand $Y$ in our models is TC intensification ($DV$), predictors $X_i$ are m environmental and TC variables, and the prediction error $\epsilon$ has a Gaussian distribution with a mean of zero and standard deviation $\sigma$. Observations of the variables, denoted by $y_i$ and $x_{ij} (i = 1, 2, \ldots, n)$, are used to estimate the model parameters.

2) SPARSE LINEAR MODEL (SLM)

Building on OLS, Tibshirani (1996) proposed a method for creating sparse linear models (SLMs). The method selects the most important predictors based on their correlation to the predictand by requiring the following constraint to Eq. (1b):

$$\sum_{j=1}^{m} |\beta_j| \leq t, \quad (2)$$

where $t$ is an adjustable “budget” parameter that constrains the regression coefficient estimates. As $t$ increases from zero to infinity, the model progressively selects variables from no variables to all variables. Equation (1b) under the constraint in Eq. (2) can be solved as

$$\begin{align*}
\hat{(\alpha, \beta)} = \argmin_{\beta} \left[ \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \alpha - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right],
\end{align*} \quad (3)$$

where $\lambda$ is a nonnegative regularization parameter that is related to $t$. Equation (3) is the Lagrangian function of Eq. (1b) with constraint Eq. (2), known as the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996).
The orange dashed line shows the value of \( \lambda_{1se} \), and the model for each of a range of validation schemes. Specifically, the data are randomly accuracy and the number of predictors. We use a cross-necessary coefficients to zero. The value of the regula-
important variables simultaneously by shrinking un-
standard deviation of the MSE for each
is repeated a number (10) of times to obtain the mean and
error (MSE) calculated for the validation set; this analysis
developed using the training set with the mean square
l value is chosen to be slightly greater than the value of

![Fig. 1. MSEs for different values of the regulation parameter \( \lambda \)
for the SLM using the NA data with the full variable selection pool.
The mean (red dots) and standard error (gray thin curves) for the
MSE for each value of \( \lambda \) are calculated via cross validation. The
green dashed line shows the value of \( \lambda \) that gives a minimum error.
The orange dashed line shows the \( \lambda \) that gives an error one standard
error away from the minimum error (\( \lambda_{1se} \)), and the blue dashed line
shows the two-standard-error line (\( \lambda_{2se} \)). The number of selected
variables for each value of \( \lambda \) is shown at the top of the plot.

Instead of performing significance tests to add each individual predictor as in the SHIPS model construction (DeMaria and Kaplan 1999), LASSO selects the most important variables simultaneously by shrinking unnecessary coefficients to zero. The value of the regularization parameter \( \lambda \) is chosen to balance the model’s accuracy and the number of predictors. We use a cross-validation scheme. Specifically, the data are randomly partitioned into a training set (90%) and a validation set (10%), and the model for each of a range of \( \lambda \) values is developed using the training set with the mean square error (MSE) calculated for the validation set; this analysis is repeated a number (10) of times to obtain the mean and standard deviation of the MSE for each \( \lambda \) value. Then, a \( \lambda \) value is chosen to be slightly greater than the value of \( \lambda_{min} \), which gives the minimum mean MSE. Common choices of \( \lambda \) are one standard error from \( \lambda_{min} \), denoted as \( \lambda_{1se} \) (the MSE corresponding to \( \lambda_{1se} \) equals the minimum MSE plus its one standard deviation), and two standard errors from \( \lambda_{min} \), denoted as \( \lambda_{2se} \). To illustrate, Fig. 1 shows an example cross-validation plot of the MSE as a function of \( \lambda \), where the dashed green line shows \( \lambda_{min} \); the dashed orange line shows \( \lambda_{1se} \), and the dashed blue line shows \( \lambda_{2se} \); the number of parameters selected for each \( \lambda \) value is also displayed. In addition, it is noted that regression coefficient estimates from LASSO may include some of the bias due to the shrinkage applied during variable selection (Zhang and Huang 2008). Thus, to obtain SLMs, we apply OLS regression on the variables selected by LASSO.

3) SPARSE GENERALIZED ADDITIVE MODEL (SGAM)

Considering that the effects of the predictors for TC intensification may be nonlinear, we use sparse generalized additive models (SGAMs; Huang et al. 2010) to identify and characterize nonlinear effects, while still imposing a constraint to limit the number of predictors. To model the nonlinearity effects, consider the non-parametric additive model:

\[
Y = \alpha + \sum_{j=1}^{m} f_j(x_j) + \varepsilon, \varepsilon \sim N(0, \sigma^2),
\]

where \( f_j \) are unknown nonlinear functions. Generalized additive models (GAMs) provide a useful extension to linear models, as they allow nonlinear transforms of predictors while still retaining much of their simplicity. In the above model, if we represent each function \( f_j \) as a linear combination of basis functions, the resulting model can then be fit by simple least squares, and a sparse model can be achieved by imposing a constraint on the basis functions. More specifically, in our analysis, we expand each \( f_j \) using B splines with degree 3:

\[
f_j(x) = \mathbf{\beta}_j \cdot \mathbf{B}(x) = \beta_{j1} B_1(x) + \beta_{j2} B_2(x) + \beta_{j3} B_3(x).
\]

Then, we have

\[
(\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \ldots, \hat{\beta}_{m1}, \hat{\beta}_{m2}, \hat{\beta}_{m3})
\]

\[
= \arg\min_{(\alpha, \mathbf{\beta})} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - \alpha - \sum_{j=1}^{m} \sum_{l=1}^{3} \beta_{jl} B_l(x_{ij}) \right] ^2 + \lambda \sum_{j=1}^{m} \| \mathbf{\beta}_j \|_2 \right\},
\]

where the constraint is imposed on the L-2 norm of the regression coefficients. Here, we apply a group LASSO penalty (Yuan and Lin 2006; Huang et al. 2010; Simon et al. 2013) on the coefficients—a generalized LASSO that sets coefficients of the basis functions for the same variable to zero simultaneously. As for SLM, \( \lambda \) is chosen based on the trade-off between fewer variables selected and lower MSE using the cross-validation scheme. The resulting SGAM has the most significant predictors selected and models the predictand as a sum of nonlinear functions of the selected predictors.

4) FINITE MIXTURE REGRESSION (FMR) LASSO

The assumption of homogeneity that requires the regression relationships to be the same for all TC intensification observations may be inadequate. Thus,
substantial prediction improvements may be possible by incorporating a heterogeneous structure to the model. The finite mixture regression (FMR) model is a widely used approach for analyzing heterogeneity data. In this model, each observation is considered to be a random sample generated by one of its k components (groups/clusters), while the identity of the generating component is not observed. In a Gaussian setting, each of the k components is a linear regression, as in Eq. (1a); then, the probability distribution of the predictand \(Y\), given the predictors \(X_j\), is

\[
f_{Y|X_1, X_2, \ldots, X_m} = \sum_{r=1}^{k} \pi_r \frac{1}{\sqrt{2\pi\sigma_r}} \exp \left[ -\frac{(y - \alpha_r - \sum_{j=1}^{m} \beta_{rj}X_j)^2}{2\sigma^2_r} \right],
\]

where \(\pi_r\) is the mixture weight (proportion of observations), \(\alpha_r\) and \(\beta_{rj}\) are regression coefficients, and \(\sigma_r\) is the standard deviation, for each group \(r\). These parameters can be estimated from data using the generalized expectation–maximization (EM) algorithm (McLachlan and Krishnan 1997), a widely used technique for parameter estimation in mixture models. In addition, the EM algorithm estimates the probability that each observation \((x_i, y_i)\) is generated from each group. Thus, FMR has also been widely used for clustering the data; for example, Gaffney et al. (2007) and Camargo et al. (2007) apply FMR analysis to group storm trajectories, based on the estimated membership probabilities, into different clusters that depict the characteristic tracks (as functions of travel time) of the storms.

In our application to model TC intensity as a function of environmental and TC variables, it is ideal to add a constraint/penalty for the variable selection to FMR. We apply the L1-penalized finite mixture of the regression model (also known as FMR LASSO) proposed by Städler et al. (2010). The method applies an efficient generalized EM algorithm to obtain the sparse linear regression function for each group. The algorithm is iterative and starts by assigning observations to random groups and maximizing the following penalized log-likelihood function with respect to the group assignment:

\[
\frac{1}{n} l_{\text{pen}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \sum_{r=1}^{k} \pi_r \frac{1}{\sqrt{2\pi\sigma_r}} \exp \left[ -\frac{(y_i - \alpha_r - \sum_{j=1}^{m} \beta_{rj}x_{ij})^2}{2\sigma^2_r} \right] \right\} - \lambda \sum_{r=1}^{k} \sum_{j=1}^{m} \frac{|\beta_{rj}|}{\sigma_r},
\]

where \(\theta\) represents the collection of model parameters \((\pi_r, \alpha_r, \beta_{rj}, \sigma_r)\) and the constraint on the L-1 norm of the regression coefficients is applied so that the variable selection is performed to obtain a sparse model within each group. Then, the model error is calculated and used to find the optimal group assignment for the observations. Similar to SLM and SGAM, \(\lambda\) is chosen with the cross-validation scheme to balance the model’s accuracy and the number of predictors.

FMR LASSO is a mixture of linear models; we also develop a mixture of nonlinear models called mixture GAM (MGAM) LASSO, which combines the merits of mixture models and SGAM. In the first step, FMR LASSO is applied to group the data and select important variables for each cluster. In the second step, we apply GAM to variables selected for each cluster. As a result, a nonlinear regression function is obtained for each cluster. Similarly, we can apply a nonparametric model, as described in the next subsection, to variables selected for each cluster to obtain a nonparametric model for each cluster; this model is denoted as mixture nonparametric (MNP) LASSO.

FMR models provide an advanced tool for clustering and investigating the heterogeneity of the data. However, we note that, unlike the homogenous models, which can be directly used for prediction, making predictions with an FMR model requires assigning a new observation of predictors \((X)\) to a group or, alternatively, assigning group weights to the new observation. This assignment can introduce further, possibly significant, uncertainties into the prediction. We will return to this point in section 4, after showing the regression analysis results.

5) NONPARAMETRIC (NP) REGRESSION

In addition to the parametric regression methods introduced above, nonparametric (NP) multivariate regression is also applied. Instead of fitting variables of a predetermined regression model, NP regression relaxes the model assumptions and hence has the ability to capture more specific structure in the dataset. In this study we apply the kernel regression. The predictand \(Y\) can be estimated from the predictor \(X\) by the first-order kernel regression:
In Eq. (7), \( K_h(x) = (1/h)K(x/h) \), where \( K(\cdot) \) is the so-called kernel and \( h \) is the bandwidth, and we use the Gaussian kernel \( K(x) = (1/(\sqrt{2\pi})) \exp(-x^2/2) \). The bandwidth \( h \) is a tuning parameter balancing the bias–variance trade-off: smaller \( h \) leads to better in-sample fit but may cause overfitting; larger \( h \) yields a smoother estimation but may sacrifice accuracy of estimation. In practice, one can use a data-driven approach such as cross validation to choose the optimal bandwidth for nonparametric kernel regression. Similarly, the prediction for \( Y \) from the second-order kernel regression with two covariates is

\[
\hat{Y} = f(X_1, X_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} y_i K_{h_1}(X_1 - x_{1i}) K_{h_2}(X_2 - x_{2i})}{\sum_{i=1}^{n} K_{h_1}(X_1 - x_{1i}) K_{h_2}(X_2 - x_{2i})},
\]

where \( h_1 \) and \( h_2 \) are the bandwidths in the two covariate directions, respectively.

In our applications, more than two predictors may be considered. However, it is noted that kernel regression with more than two covariates often performs poorly with relatively large errors (Fan and Gijbels 1996, chapter 7). Thus, we apply kernel regression with model aggregation. Specifically, we model the response \( Y \) as a linear combination of first- and second-order kernel regression functions of \( m \) predictors \( X_j \) \((m > 2)\) as

\[
Y = \alpha + \sum_{j=1}^{m} \beta_j f_j(X_j) + \sum_{p \neq q} \beta_{pq} f_{pq}(X_p, X_q) + \varepsilon , \varepsilon \sim N(0, \sigma^2),
\]

where \( f_j \) are first-order kernel regression functions fitted for \( X_j \), as in Eq. (7), and \( f_{pq} \) are second-order kernel regression functions fitted for \( X_p \) and \( X_q \), a combination of any two different covariates in the model, as in Eq. (8). By applying lower-order kernel regression and aggregating the fitted functions with an additional linear model, we can reduce the prediction error for higher-order analysis and at the same time preserve the advantage of nonparametric regression.

The goodness of fit of the NP regression is, as for other models, evaluated by the \( R \)-squared \((R^2) \) value, which reflects the percentage of variation in the predictand variable explained by the predictors. All statistical analyses in this study are calculated using the R statistical software interface. The glmnet package (Friedman et al. 2010) is used for the SLM analysis, the gglasso (Yang and Zou 2015) and gam (Hastie and Tibshirani 1990) packages are used for the SGAM and GAM analyses, the fmrlasso package (Städler et al. 2010) is used for the FMR LASSO analyses, and the np package (Hayfield and Racine 2008) is used for the nonparametric analysis.

3. Results

3.1 Variable selection and linear–nonlinear regression

The OLS [Eq. (1)], SLM [Eq. (3)], and SGAM [Eq. (4)] methods are performed on the observations of \( DV \) and 13 potential predictor variables for NA and WNP. Table 2 shows the selected variables using each method. The OLS columns list all the variables for linear regression as a baseline for comparison. The SLM columns list the variables selected by LASSO for linear regression. The SGAM columns list the variables selected by group LASSO for nonlinear regression. The variables appear in the order of significance as determined by their contribution to the explained variance of the predictand, and the sign of the coefficient for each variable is also shown. The relative order of importance among the predictors is similar across models. Storm-persistent variables (i.e., current intensity \( V \) and last-step intensification \( DV \)) are most significant, as expected, especially for the short-time prediction. Among the environmental variables, MPI, SHR, RHHI, and U200 are most important; Z850 is also relatively important. Specifically, intensification is favored if the storm has been intensifying, has a relatively low intensity, and is surrounded by an environment that has high MPI, low wind shear, high upper-level relative humidity, high-level zonal wind more easterly than normal, and high low-level vorticity. These observations are consistent with previous studies; in particular, DeMaria and Kaplan (1999) point out that 200-hPa winds are more easterly than normal equatorward of an upper-level ridge, which is a favorable region for intensification. Another large-scale forcing term, D200, is relatively less important for the 6-h intensification considered here, consistent with previous observations that this variable is only significant for longer prediction windows (72 h; Lee et al. 2015). The variable T200 is expected to have an inverse relationship with intensification because a lower upper-level atmospheric temperature may increase the efficiency of convection by providing a larger temperature difference from the surface. However, T200 is not significant here, as the upper-level temperature is already applied in the calculation of MPI.

A few variable predictors have an unexpected sign for their coefficient estimate according to the preestablished...
Table 2. OLS, SLM, and SGAM analyses for modeling intensification (DV) for NA and WNP. Selected variables are listed in their order of statistical significance, followed by the sign of their coefficient (for linear models). The bottom row shows the $R^2$ value for each model.

<table>
<thead>
<tr>
<th>North Atlantic</th>
<th>Western North Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OLS</strong></td>
<td><strong>SLM</strong></td>
</tr>
<tr>
<td>1</td>
<td>$DV_r$</td>
</tr>
<tr>
<td>2</td>
<td>SHR</td>
</tr>
<tr>
<td>3</td>
<td>$V$</td>
</tr>
<tr>
<td>4</td>
<td>U200</td>
</tr>
<tr>
<td>5</td>
<td>LSHR</td>
</tr>
<tr>
<td>6</td>
<td>Z850</td>
</tr>
<tr>
<td>7</td>
<td>MPI</td>
</tr>
<tr>
<td>8</td>
<td>D200</td>
</tr>
<tr>
<td>9</td>
<td>USHR</td>
</tr>
<tr>
<td>10</td>
<td>T200</td>
</tr>
<tr>
<td>11</td>
<td>TSPD</td>
</tr>
<tr>
<td>12</td>
<td>RHHI</td>
</tr>
<tr>
<td>13</td>
<td>JDAY</td>
</tr>
</tbody>
</table>

$R^2$ | 0.26 | 0.24 | 0.26 | 0.30 | 0.30 | 0.31

physical understanding. While the sign of the coefficient helps with the interpretability of the model, it does not necessarily mean that it is incorrect when unexpected signs appear in a multivariate regression setting, as also noted for the SHIPS and STIPS models (DeMaria and Kaplan 1999; Knaff et al. 2005). Since some predictors are interrelated, an unexpected sign may indicate compensation for the contribution of other variables. For example, SHR is expected to have an inverse relationship with intensification, since higher values of vertical wind shear inhibit convection. LSHR has a positive sign in the OLS models, as it statistically compensates for the negative effects of SHR and USHR. This statistical result is consistent with the theoretical argument of DeMaria (1996) that, as they have smaller Rossby penetration depth, low-latitude storms are more sensitive to vertical shear than higher-latitude storms. TSPD is shown to have an inverse relationship to intensification, likely because fast translational speed can lead to more significant asymmetry in the cyclone structure and inhibit intensification (e.g., Chen et al. 2006). However, previous studies for longer-time-interval analysis (>12 h) showed a positive relationship between TSPD and intensification (Lee et al. 2015), which may be explained by the fact that slow movement may lead to cold-water upwelling near the storm center and prevent intensification (e.g., Geisler 1970).

The $R^2$ values for the OLS model with 13 predictors are about 0.26 for NA and 0.3 for WNP. (As a comparison, the $R^2$ values for the OLS model with only the current intensity and last-step intensification as predictors are about 0.24 for NA and 0.22 for WNP.) These values for 6-h predictions are lower than those for 12-h predictions in the SHIPS (0.36 for NA; DeMaria and Kaplan 1999) and STIPS (0.4 for WNP; Knaff et al. 2005) models, as $R^2$ tends to increase with increasing forecast window. Compared to the OLS method, the SLM approach succeeds in significantly reducing the number of predictors (from 13 to 5 for NA and to 8 for WNP) needed to retain similar $R^2$ (0.24 for NA and 0.3 for WNP). When the nonlinear relationship is applied with the SGAM approach, more variables are selected (8 for NA and 11 for WNP), and the $R^2$ values are slightly increased (0.26 for NA and 0.31 for WNP).

b. Mixture modeling

Mixture modeling using FMR LASSO [Eq. (6)] is performed for the $DV$ predictand to capture the subpopulation heterogeneity and enhance the statistical modeling. We perform the analysis with three groups ($k = 3$) for both NA and WNP using the 13 potential predictor variables. The three groups that are defined based purely on statistical properties of the predictand and predictor variables present distinct physical features. As shown in Fig. 2, the first group is featured with little intensification or $DV$ values centering around zero (“static,” in red), the second group with moderate intensification or $DV$ taking relatively moderate positive and negative values (“normal,” in green), and the third group with rapid intensification/deintensification or large $DV$ values in both the positive and negative directions (“extreme,” in blue). Note that the groups, especially the normal and extreme, have some areas of overlap, so it is not sufficient to define them solely by thresholds of $DV$. These three regimes of intensification also appear clearly as three phases in intensification during the lifetime of individual storms, as shown by two examples in Fig. 2. The statistically classified extreme intensification indicates “rapid intensification,” although the latter is often
specifically defined for the positive increase of intensity over a longer time interval [e.g., at least 30-kt (15.4 m s\(^{-1}\)) increase in the maximum sustained wind within 24 h or less; Kaplan and DeMaria (2003)].

Table 3 lists the variables selected and regression statistics in the FMR LASSO analyses. For each group, the predictors that were selected appear in the order of statistical significance. Below the predictor lists are the statistics associated with each group. The weight for each group is \(\pi_r\) in Eq. (6), calculated as the proportion of observations from the dataset that are optimally assigned to each group \(r\). The mean (\(\mu_r\)) and standard deviation (\(\sigma_r\)) of \(DV\) are calculated for each group using the observations in each group. The static group has a mean close to zero with a small standard deviation, the normal group has a moderately higher standard deviation, and the extreme group has the largest standard deviation, as also reflected by the \(DV\) observations colored by group in Fig. 2. In assessing the statistical performance of the models, \(R^2\) values are calculated for each group separately, and the total \(R^2\) considers all groups together. Compared to the previous analyses assuming a homogenous data structure (Table 2), FMR analysis increases the \(R^2\) values significantly (0.41 for NA and 0.49 for WNP) while using fewer predictor variables (five for NA and seven for WNP). This result indicates that TC intensification is a heterogeneous mixture with varying dependencies on the TC and environmental variables. Applying GAM to account for nonlinearity in the mixture model (MGAM) slightly increases the total \(R^2\) (0.43 for NA and 0.53 for WNP), as shown also in Table 3.

In addition, compared to the homogenous models (Table 2), improvement is seen in the mixture models in matching with the physical understanding, or coefficient sign interpretability, of the relationship between the variables and intensification (Table 3). This may be due
to the fact that fewer (possibly correlated) variables are selected, but it may also indicate that some physical understanding of predictors may apply only to certain groups or phases of the TC intensification. In addition to the persistent variables, MPI and SHR are selected for both basins. The variable U200 is shown to be significant for NA, while RHHI is important for WNP. It is noted that the static group models DV for NA as a pure Gaussian random variable and for WNP using only TSPD as a predictor, indicating that a TC that maintains a constant intensity largely ignores the environment represented by the applied predictors (however, the $R^2$ value for the group is quite low). This result may also reflect the limitations of the data; the intensity values from the best-track data are rounded to the nearest 5 kt (2.58 m s$^{-1}$), and thus some of the data points in the static group may be static only because of this rounding. After this group is separated out, the normal and also the extreme groups show much improved statistical relationships between the 6-h intensity change and the selected small number of storm and environmental predictors (compared to Table 2). However, it is important to note again that in calculating these $R^2$ values, the correct group classification is assumed to assess the performance of the model in terms of capturing the statistical relationship between the predictand and predictor variables; to apply the model for prediction, one needs to first predict the group membership, which can induce significant uncertainty (see section 4).

c. Testing ventilation index (VI)

Instead of using many environmental variables, VI and TC intensity variables are explicitly selected as predictors to construct models for DV using multiple methods. In the dimensional analysis, we model $DV$ using VI, $V$, and $DV_p$. In both the OLS and GAM models, the $R^2$ values are about 0.25 for NA and 0.23 for WNP. These $R^2$ values are lower than those of the corresponding models using more predictors, especially for WNP, as shown in Table 2.

Next, we perform FMR, MGAM, and MNP analyses. To better visualize these mixture analysis results, we show the estimated function for each variable that contributes to the predictand for each group in Figs. 3 and 4 for NA and WNP, respectively. The red solid line shows the linear function from the FMR model, the solid blue curve gives the nonlinear function from the MGAM model, and the black solid curve represents the nonparametric function from the MNP model, for each predictor in each group. For instance, the top-left panel in Fig. 3 is the function $f(VI)$, which, for each value of VI, gives a value that contributes to the predictand $DV$ in the first group in the NA analysis. In most cases, the linear function appears to be a simplified trend of the nonlinear and nonparametric functions, which can also capture local variations. Large uncertainties and differences in the models are apparent mainly in the regions of sparse observations (see gray hash marks). Consistent with the VI theory (Tang and Emanuel 2012), VI has generally a negative correlation with intensification in all groups for both NA and WNP. Previous intensification ($DV_p$) is positively correlated with intensification, as expected. Although the current intensity $V$ is expected to negatively correlate with intensification, exceptions occur for the static group for WNP, likely because of statistical compensation effects. The signs of the coefficients of the linear models are also shown in Table 4, as are the main statistics.

<table>
<thead>
<tr>
<th>Cluster ($r$)</th>
<th>North Atlantic</th>
<th>Western North Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 static</td>
<td>2 normal</td>
</tr>
<tr>
<td>Ordered selected variables</td>
<td>$DV_p$ + $V$ -</td>
<td>TSPD - $DV_p$ + $V$ -</td>
</tr>
<tr>
<td>Weight</td>
<td>0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.01</td>
<td>2.27</td>
</tr>
<tr>
<td>Std dev</td>
<td>0.45</td>
<td>5.70</td>
</tr>
<tr>
<td>FMR $R^2$</td>
<td>0.00</td>
<td>0.47</td>
</tr>
<tr>
<td>FMR total $R^2$</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>MGAM $R^2$</td>
<td>0.00</td>
<td>0.47</td>
</tr>
<tr>
<td>MGAM total $R^2$</td>
<td>0.43</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Mixture LASSO analyses for modeling intensification ($DV$) using $k = 3$ groups for NA and WNP. Variables selected (from the 13 variables in Table 2) are listed in their order of statistical significance, followed by the sign of their coefficient (for linear models). The group weight and mean and standard deviation of intensification for each group are shown. The $R^2$ values for each group and for the total are shown for each of two models (FMR and MGAM).
In this case using VI as the only environmental predictor, the $R^2$ values of the FMR model (0.41 for NA and 0.46 for WNP) and of the MGAM model (0.44 for NA and 0.52 for WNP) are similar to those using more environmental variables (see Table 3), indicating that VI is a dominate variable for storm intensification. However, considering that the $R^2$ values are reduced in the OLS and GAM models when using VI, this result indicates that the heterogeneous feature is more apparent in the data when using VI to predict TC intensification. The MNP modeling, which accounts for local variability, further increases the $R^2$ values (0.49 for NA and 0.58 for WNP).

In the nondimensional analysis, we model $d(V/\text{MPI})$ using VI, $V/\text{MPI}$, and $d(V/\text{MPI})_p$, following the modeling framework of Tang and Emanuel (2012). Generally, a positive $d(V/\text{MPI})$ means that a storm is intensifying since MPI is relatively slowly varying in time and space. Similar to the dimensional analysis results, the simple OLS and GAM models do not perform as well as those using other environmental variables, with $R^2$ values of only about 0.24 for both NA and WNP.

Mixture regression analyses are performed on the nondimensional data. Table 5 shows the results for the three new groups determined for the nondimensional predictand $d(V/\text{MPI})$ for each ocean basin. Given the nondimensional form, the three groups are no longer clearly distinguished as the static, normal, and extreme clusters; the groups have quite close values of the standard deviation of the predictand for NA. Most signs of coefficients are as expected; namely, $d(V/\text{MPI})$ is associated positively with $d(V/\text{MPI})_p$ and negatively with $V/\text{MPI}$ and VI. But VI shows positive correlation with intensification for groups 2 and 3 for WNP, and $V/\text{MPI}$ shows positive correlation with intensification for group 3 for NA, as a result of statistical compensation effects and possibly also other effects such as ocean mixing that we do not consider here. It is noted that the total $R^2$ values are significantly increased (0.7–0.74 for NA and 0.49–0.62 for WNP) in the nondimensional analysis with VI, compared to the dimensional analysis (see Table 4), indicating that the VI effect may be better represented in a nondimensional model.

d. Comparing the predictability of significant variables using nonparametric analysis

It is thus interesting to investigate how the specific nonlinear function of TC intensification derived theoretically based on the VI theory (Tang and Emanuel 2012)
would fit with the real data. Tang and Emanuel (2012, their Fig. 7) show an excellent comparison of the observed and theoretically predicted 24-h normalized intensification \( [d(V/MPI)] \) as a function of VI and normalized intensity \( (V/MPI) \). It is noted that, to capture the main features in the comparison between the observation and theory, the observational data in their figure are binned (i.e., average value of the predictand is taken within the bin grid of the predictor space). Here, we plot a similar binned figure for 6-h intensification, on which we focus in this study, for the NA data (Fig. 5a). A pattern similar to that predicted by the theoretical function is obtained [note that the time interval over which the intensification occurs is ambiguous in the theory; Tang and Emanuel (2012)]. Statistical regression using the functional form from the theory on the binned data gives high \( R^2 \) values. In fact, even a simple linear model gives a similarly high \( R^2 \) value of 0.6 for the 6-h intensification data (and 0.64 for 24-h intensification data). However, this result is heavily dependent on the choice of the bin width. The \( R^2 \) values could drop to less than 0.1 if the bin width becomes very small. This possibility exists because the original data of

Table 4. As in Table 3, but here the VI and TC intensity variables are explicitly selected as predictors and MNP analysis results are also shown.

<table>
<thead>
<tr>
<th>Cluster ((r))</th>
<th>North Atlantic</th>
<th>Western North Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered variables</td>
<td>1 static</td>
<td>2 normal</td>
</tr>
<tr>
<td>VI</td>
<td>(-)</td>
<td>(+)</td>
</tr>
<tr>
<td>(DV_P)</td>
<td>(+)</td>
<td>VI</td>
</tr>
<tr>
<td>V</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Weight</td>
<td>0.32</td>
<td>0.57</td>
</tr>
<tr>
<td>Mean</td>
<td>(-0.01)</td>
<td>2.21</td>
</tr>
<tr>
<td>Std dev</td>
<td>0.48</td>
<td>5.63</td>
</tr>
<tr>
<td>FMR (R^2)</td>
<td>0.01</td>
<td>0.54</td>
</tr>
<tr>
<td>FMR total (R^2)</td>
<td>0.01</td>
<td>0.41</td>
</tr>
<tr>
<td>MGAM (R^2)</td>
<td>0.03</td>
<td>0.55</td>
</tr>
<tr>
<td>MGAM total (R^2)</td>
<td>0.03</td>
<td>0.44</td>
</tr>
<tr>
<td>MNP (R^2)</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td>MNP total (R^2)</td>
<td>0.16</td>
<td>0.49</td>
</tr>
</tbody>
</table>
6-h intensification, as shown by the scatterplot in Fig. 5b, are very "noisy." Without the averaging/smoothing, the noise even dominates and \(d(V/\text{MPI})\) can no longer be explained by a simple linear function or even the theoretical nonlinear function of VI and \(V/\text{MPI}\). When we apply the nonparametric modeling (without grouping) to the 6-h normalized intensification \([d(V/\text{MPI})]\) as a function of VI, normalized intensity \((V/\text{MPI})\), as well as previous intensification \([d(V/\text{MPI})_p]\), we obtained \(R^2\) values of 0.31 for NA and, curiously, as high as 0.54 for WNP.

To further compare the predictability of VI, its components, and other previously identified significant variables, we perform a series of nonparametric modeling analyses (without grouping and in dimensional forms) of 6-h intensification with these predictors, as shown in Table 6. (OLS and GMA models are also applied in parallel, but they perform less well than NP for all cases and thus are not shown.) First, compared to the non-dimensional modeling discussed above, the dimensional VI model (model 2) performs similarly (in terms of the \(R^2\) value) for NA but less well for WNP. VI is shown to perform better (model 2) than MPI, SHR, or ED (models 3–5), which are the three components of VI. The model with VI (model 2) is also comparable to the model with all three component variables of VI (model 9). Applying the three component variables provides a statistically better model than applying VI, because, as mentioned above, the real data are noisy, and applying the three variables separately provides a greater degree of freedom for statistical fitting. We also test RHHI in place of ED. RHHI performs less well than ED when applied as the sole environmental variable or with MPI or SHR as the other environmental variable (models 10–12 compared to models 3 and 7–8). When RHHI is applied in place of ED with both MPI and SHR, the results are similar (\(R^2\) slightly lower for NA and slightly higher for WNP; model 13 compared to model 9), as the significance of ED/RHHI is reduced when both MPI and SHR are applied. Finally, we add U200, which

<table>
<thead>
<tr>
<th>Cluster (r)</th>
<th>North Atlantic</th>
<th>Western North Pacific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ordered variables</td>
<td>(d(V/\text{MPI})_p)</td>
<td>+</td>
</tr>
<tr>
<td>VI</td>
<td>-</td>
<td>VI</td>
</tr>
<tr>
<td>Weight</td>
<td>0.62</td>
<td>0.21</td>
</tr>
<tr>
<td>Mean</td>
<td>0.01</td>
<td>-0.05</td>
</tr>
<tr>
<td>Std dev</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>FMR (R^2)</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td>FMR total (R^2)</td>
<td>0.70</td>
<td>0.49</td>
</tr>
<tr>
<td>MGAM (R^2)</td>
<td>0.43</td>
<td>0.47</td>
</tr>
<tr>
<td>MGAM total (R^2)</td>
<td>0.71</td>
<td>0.49</td>
</tr>
<tr>
<td>MNP (R^2)</td>
<td>0.46</td>
<td>0.66</td>
</tr>
<tr>
<td>MNP total (R^2)</td>
<td>0.74</td>
<td>0.62</td>
</tr>
</tbody>
</table>

*Fig. 5. Comparison of (a) bin-averaged and (b) scatterplot of intensification \([d(V/\text{MPI})]\) as a function of \(V/\text{MPI}\) and VI for NA. In (a), the VI axis is divided logarithmically into 30 grids between 0.001 and 1, and the \(V/\text{MPI}\) axis is divided evenly into 24 grids between 0 and 1.2. Bins with sample sizes less than three are omitted.*
may not affect TC intensification in a direct, physical way as do the other discussed variables but has been shown to be statistically significant. Adding U200 significantly improves $R^2$ in all models for both NA and WNP (models 14–20); however, the role U200 plays in controlling TC intensification still calls for further investigation. The model with six predictors ($V, DV_p, MPI, ED, SHR,$ and U200) gives the highest $R^2$ values (about 0.37 for NA and 0.36 for WNP; model 20) among these nonparametric models.

4. Discussion

The mixture regression analyses reveal the heterogeneous properties of the data and provide models that better explain the statistical relationship between storm intensification and the environmental and storm predictors. However, the probability of belonging to a group, or the proportion $\pi_r$, does not depend on the predictors $X$ in the current FMR or FMR LASSO, so the mixture model is not readily a predictive model (i.e., the heterogeneity cannot be directly predicted). To add the predictive capability to the mixture models, therefore, we build classification models using classic methods such as multinomial logistic regression and probit regression. Such a model determines the probability of a new observation $X$ belonging to each of the groups as a function of $X$. Thus, for a new observation, the prediction (of storm intensification) can be taken according to the group model with the highest probability or as a probability-weighted average of the group model predictions. For evaluation, a predictive $R^2$ can be calculated using the predicted group assignments. The value of the predictive $R^2$, which accounts for the uncertainty in assigning the group to a new observation, should be lower than that of the standard $R^2$, which is calculated given the optimized group assignment.

As a result, the predictive $R^2$ is greatly reduced (to below 0.3), compared to the standard $R^2$ (0.41–0.74), for all mixture models presented in section 3. To investigate where the most significant uncertainties are introduced, we employ a loss function to quantify the behavior of the classification model, or the accuracy of the group re-assignment. The loss function of observation $i$ is designed to be $\text{LOSS}_i = \sum_r (P_{r,i} - 1, i)^2$, where $P_{r,i}$ is the probability based on the classification model that the observation $i$ belongs to group $r$ and $1, i$ is the indicator function equal to 1 if $i$ belongs to group $r$ or 0 if not, according to the actual (optimized) group assignment. Thus, the lower the probability assigned to the right group based on the classification model, the larger the loss. Figure 6 shows an example of the calculated loss for one mixture model with logistic regression; similar results are obtained for other mixture models in section 3 and for using both the logistic and multinomial regression methods.

The calculated loss function indicates that the group assignment has the highest accuracy for the normal group 2 observations, followed by the static group 1 (which is affected by data accuracy as discussed before). The extreme group 3 observations are often assigned to an incorrect group, which is the main reason that the
overall predictive $R^2$ is greatly reduced. This result indicates that the selected predictors have limited predictability for the occurrence of rapid intensification. One reason for this limitation may be that oceanic variables are not explicitly included in this study focusing on the large-scale environmental factors. However, Kaplan et al. (2010) also obtain only low-to-modest predictability of RI occurrence even when the ocean heat content and other satellite-derived variables were added as additional predictors. Kaplan et al. (2010) highlight the difficulty of predicting RI occurrence and suggest further examining other predictors, especially those related to the inner-core process. Various new predictors related to oceans, inner-core processes, and storm structures are continuously being explored and added to the operational forecasting of RI (Kaplan et al. 2015).

For the purpose of prediction, the best model we obtained is the nonparametric model with six predictors ($V, DVP, MPI, ED, SHR, \text{ and } U200$) for NA and the nonparametric VI model in the nondimensional form for WNP. The nonparametric regression, which often requires larger sample sizes than parametric regression, actually accounts for the heterogeneity of the data. In the kernel regression, the prediction $Y$ for a new observation of predictors $X$ is essentially a weighted average of the $y$ observations in the dataset with the weights depending on the new observation [Eqs. (7) and (8)]. As the predictors applied in the NP analysis are similar to those in the mixture models, it is likely that the extreme intensification is not well modeled either in the NP analysis. Here, oceanic and inner-core-process variables should also be explored in future studies.

In addition, the predictability of these models may be improved by using longer-time-interval or “smoothed” intensity change predictands. For example, the intensity change as well as the likelihood of rapid intensification over a 24-h interval may be predicted better statistically, as shown by previous studies (DeMaria and Kaplan 1999; Knaff et al. 2005; Lee et al. 2015; Kaplan et al. 2015). Also relevant is the “intensification tendency,” which is calculated as the linear regression coefficient of the wind intensity over six successive 6-hourly steps starting from the current time (Lloyd and Vecchi 2011). These longer-time-window intensity changes are less affected by data accuracy and precision, and by smoothing out the (possibly physical) noise, they may reflect better the temporal scale of the TC intensification in response to the atmospheric and oceanic conditions. In forecasting and risk analysis, however, additional models may be needed to predict shorter-time-interval intensity changes from the longer-time-interval estimates.

5. Summary

In this study, we have explored the dependence of TC intensification on the environment through various statistical modeling approaches. The NA and WNP TC observations from the IBTrACS WMO archive and surrounding environmental variables from the ERA-Interim over 1979–2014 were used. The developed statistical models estimate the 6-h intensity change (predictand) given the environmental and storm variables (predictors).

First, the simplest multiple linear regression model using 13 TC and environmental variables informed by the operational SHIPS and STIPS models obtained statistical $R^2$ values of about 0.26 for NA and 0.3 for WNP. The variable selection technique LASSO was applied to identify the critical predictors simultaneously; the analysis shows that intensification is favored if the storm has been intensifying, has a low intensity, and is surrounded by an environment that has high MPI, low wind shear, high upper-level relative humidity, and high-level zonal winds that are more easterly than normal, consistent with previous studies. The SLM model, which applies linear regression to LASSO-selected variables, succeeds in significantly reducing the number of predictors (from 13 to 5 for NA and to 8 for WNP) needed to retain similar $R^2$ results (0.24 for NA and 0.3 for WNP). When the nonlinear relationship is applied with the SGAM approach, more variables (8 for NA and 11 for WNP) are selected and the $R^2$ values are slightly increased (0.26 for NA and 0.31 for WNP).

Second, mixture LASSO modeling was performed to capture the subpopulation heterogeneity and enhance the statistical modeling. The modeling separates the data into subgroups based on the statistical properties of the predictand and predictor variables, selects critical predictors for each group, and develops the regression model for each group. The statistical analysis identified...
three intensification regimes: static, normal, and extreme, reflecting observed physical features of intensification during a storm’s life cycle. The mixture modeling significantly increased the $R^2$ values (0.41 with linear modeling and 0.43 with nonlinear modeling for NA; 0.49 with linear modeling and 0.53 with nonlinear modeling for WNP) with greatly reduced numbers of predictors (five for NA and seven for WNP). However, the mixture model is not readily a predictive model, and a classification model is needed to assign the group membership for a new observation of predictors. When a classic classification model such as the multinomial logistic regression or probit regression was added, the overall predictive $R^2$ values of the mixture models were reduced to below 0.3. The main reason for this significant reduction in the predictability of the mixture models is that the applied predictors, which are mainly large-environmental variables, are limited in their ability to predict the occurrence of extreme/rapid intensification.

Third, the progression of analyses from linear, nonlinear, to mixture modeling was applied to test VI as the environmental predictor in both dimensional and nondimensional forms. The homogenous linear and nonlinear models obtained relatively low $R^2$ values (0.25 for NA and 0.23 for WNP in dimensional analyses and 0.24 for NA and WNP in nondimensional analyses). However, the $R^2$ values were greatly improved when the mixture modeling was applied (0.41/0.7 with linear modeling, 0.44/0.71 with nonlinear modeling, and 0.49/0.74 with nonparametric modeling for NA in dimensional/nondimensional analyses: 0.46/0.49 with linear modeling, 0.52/0.49 with nonlinear modeling, and 0.58/0.62 with nonparametric modeling for WNP). These analysis results indicate that VI is a dominant environmental factor for TC intensification, more so in the nondimensional form than the dimensional form, but its effect on TC intensification is quite heterogeneous and nonlinear. However, when a classic classification model was added, the overall predictive $R^2$ values of the VI mixture models were also reduced to below 0.3.

Finally, the nonparametric method was applied to the analysis of VI, its component variables, and other variables highlighted by the variable selection analysis. We found that VI, as a theoretically based combination of three significant environmental variables (MPI, SHR, and ED), can be conveniently considered to be a single environmental variable for intensification analysis. However, applying the three component variables separately may provide a better model statistically because the real data is very noisy and applying the three variables separately provides a greater degree of freedom for statistical fitting. The simpler RHRI represents ED relatively well statistically. Adding U200 significantly improves the predictability of the statistical models. The nonparametric model with six predictors ($V, DV_p, MPI, ED, SHR,$ and U200) gives relatively high predictive $R^2$ values (0.37 for NA and 0.36 for WNP). It was also found that nonparametric modeling with VI in the nondimensional form gives an $R^2$ value as high as 0.54 for WNP. The reason behind this high statistical performance and how to apply nondimensional analysis involving other predictors requires further investigation.

This work has helped to better explain the relationship between TC intensification and the surrounding environment by exploring new ways to connect them statistically. We have come up with several viable alternatives to using the simple linear regression models that depend on a long list of environmental variables. We have narrowed down the essential predictors to a smaller pool that is more favorable for real-time forecasting and risk modeling. In contrast to Lee et al. (2015) and the SHIPS (DeMaria and Kaplan 1999) and STIPS (Knaff et al. 2005) models, we applied new variable selection techniques, employed methods to take into account the nonlinear effects of the predictors, and investigated different subgroups of data that can align with a physical understanding. Also unique to any statistical modeling of TC intensity, we have incorporated the new variable VI and carefully examined its application. Tang and Emanuel (2012) motivated its use with physics and provided supporting example cases. We tested its use in formal statistical modeling.

The predictability of the mixture and nonparametric models may be further improved by adding other predictors. In our ongoing work, ocean variables (Schade and Emanuel 1999) are being added to the predictor pool, and the effects of the ocean on other variables like MPI (Lin et al. 2013; Balaguru et al. 2015) are being incorporated. Longer-time-interval (e.g., 24-h intensity change) and “smoothed” (e.g., intensification tendency) predictands are also being investigated. The results of these efforts will be reported upon in the near future. The next step of this work is to use the models for the simulation of TC intensity. Simulations may be compared to real storms as further evaluation of comparable models. Matters of initialization, termination, overland behavior, and extratropical characteristics must be considered for simulations.

Acknowledgments. This study is supported by Grants 1514606 from the National Science Foundation and NA14OAR4320106 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations herein are those of the authors and do not necessarily reflect the views of the National Science
Foundation, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce. We thank Brian Tang for helping us with the calculation of the ventilation index and providing useful knowledge of its properties. We also thank Daniel Chavas and two anonymous reviewers for their very helpful comments.

REFERENCES


