Variable Selection for Tropical Cyclogenesis Predictive Modeling

JASPER S. WIJNANDS

School of Mathematics and Statistics, and Melbourne School of Design, University of Melbourne, Parkville, Victoria, Australia

GUOQI QIAN

School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, Australia

YURIY KULESHOV

School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, and Australian Bureau of Meteorology, Docklands, Victoria, and School of Science, Royal Melbourne Institute of Technology (RMIT) University, and Faculty of Science, Engineering and Technology, Swinburne University of Technology, Melbourne, Victoria, Australia

(Manuscript received 17 May 2016, in final form 13 August 2016)

ABSTRACT

Variable selection for short-term forecasting (up to 72 h) of tropical cyclone (TC) genesis has been investigated. IBTrACS data (1979–2014) are used to identify the genesis time and position of over 2500 TCs between 30°N and 30°S. Tracks are extended using a tropical cloud cluster (TCC) dataset, which is also used to identify over 28 000 nondeveloping TCCs. Subsequently, corresponding local environment states at various atmospheric pressure levels are retrieved from ERA-Interim data. An initial selection of potentially favorable variables for TC genesis is made based on mutual information, which forms the set of nodes for graphical model structure learning using the Peter–Clark (PC) algorithm. Structure learning identifies the variables with the strongest influence on TC genesis, while taking into account the interrelationship with other variables. Variables are ranked based on the maximum observed p value in all (conditional) independence tests of the variable with the TC genesis node. The results indicate that potential vorticity (600 hPa), relative vorticity (925 hPa), and (vector) vertical wind shear (200–700 hPa) are the highest ranked variables for forecasting up to 72 h. These are followed by the basin and zonal wind speed (200 hPa), and for very short lead-time divergence (925 hPa), air temperature (300 hPa), and average vertical velocity. Predictive modeling with logistic regression confirms the superior performance of the top-ranked variables. The presented variable ranking (methodology) can be used as a building block for the creation of genesis indices or predictive models in the future.

1. Introduction

Tropical cyclones (TCs) are extreme weather phenomena that form over warm tropical waters. When TCs approach a populated coastal area they pose risks to life, property, and the environment. Hence, skillful forecasting of TCs could mitigate risks by allowing for timely planning and preparedness measures. Therefore, TC genesis forecasting is investigated on a variety of time scales, including long-term (e.g., Emanuel 2013; Murakami et al. 2012), multiyear (e.g., Smith et al. 2010; Vecchi et al. 2013), seasonal (e.g., Vecchi et al. 2014; Wijnands et al. 2015), intraseasonal (e.g., Slade and Maloney 2013; Vitart et al. 2010), and short-term forecasts (e.g., Halperin et al. 2013; Shen et al. 2010). This study focuses on short lead times (less than 72 h).

Much research has been performed to uncover the fundamental drivers of the formation of TCs from an atmospheric science perspective. The pioneering studies by Gray (1975, 1998) suggested the following six main drivers for TC genesis: (i) the Coriolis parameter, (ii) low-level relative vorticity, (iii) vertical wind shear between 200 and 950 hPa, (iv) sea surface temperature

Corresponding author address: Yuriy Kuleshov, Australian Bureau of Meteorology, GPO Box 1289, Melbourne, VIC 3001, Australia.
E-mail: y.kuleshov@bom.gov.au

DOI: 10.1175/MWR-D-16-0166.1

© 2016 American Meteorological Society

(SST), (v) difference in equivalent potential temperature between the surface and 500 hPa, and (vi) relative humidity at 500–700 hPa. Researchers have suggested various concepts for TC genesis, such as conditional instability of the second kind (CISK) (Charney and Eliassen 1964) and wind-induced surface heat exchange (WISHE) (Emanuel 1987). Other studies have explored the role of enhanced convection from equatorial waves on the formation of TCs (e.g., Schreck et al. 2012).

With increased knowledge of the underlying physical processes, methods to forecast TC genesis received considerable attention. Several indices have been designed to give an indication of the TC genesis potential of a specific environment. For example, Gray (1975) suggested a seasonal index based on the multiplication of the six drivers described above. McBride and Zehr (1981) suggested a daily genesis potential based on the difference between relative vorticity at 900 and 200 hPa. Ward (1995) developed a genesis index based on relative vorticity at 1000 hPa, the zonal component of the vertical wind shear between 200 and 1000 hPa, and SSTs above 27.6°C. Another TC genesis parameter was proposed by DeMaria et al. (2001) and is based on the zonal component of vertical wind shear between 200 and 850 hPa, a vertical instability variable, and humidity measured over various levels in the midlevel atmosphere. Emanuel and Nolan (2004) suggested a further genesis index consisting of relative humidity at 700 hPa, absolute vorticity, potential intensity, and vector wind shear between 200 and 850 hPa. Camargo et al. (2007) applied this index using relative humidity at 600 hPa and absolute vorticity at 850 hPa.

Besides indices, several statistical prediction models have been developed to forecast TC genesis. Chand and Walsh (2011) used a probit regression model to forecast TC genesis in regions of the South Pacific Ocean (Fiji, Samoa, and Tonga). Variables included in their final model are vector wind shear between 200 and 850 hPa, relative vorticity at 850 hPa, and relative humidity between 500 and 700 hPa. Zhang et al. (2015) used a decision tree algorithm for short-term TC genesis forecasts. The selected variables in this study are relative vorticity at 800 hPa, SST, precipitation rate, divergence averaged between 500 and 1000 hPa, and air temperature anomaly at 300 hPa.

One can see that many researchers have proposed an index or developed a predictive model for TC genesis. However, the selected variables and atmospheric levels at which the variables are measured vary considerably among these models. Therefore, a thorough analysis of variable selection could be advantageous, which is what this study provides. In this research global data and mathematical techniques are used to obtain the relationships of environmental variables with TC genesis, where the interaction effects between the variables are also taken into account. This determines which variables are most appropriate to be used in such an index or predictive model. In addition, this analysis is performed for varying lead times to determine any changes in relative importance of the variables. This study, therefore, focuses on a structured approach for variable selection.

The paper is organized as follows: section 2 specifies the data used, the data processing steps, and the related assumptions. It also contains a description of the mathematical methodology used for learning associative relationships from data and an explanation of the validation setup. In section 3 the results of this study are presented. Section 4 contains a discussion and a summary is provided in section 5.

2. Data and methodology

a. Data collection and processing

The World Meteorological Organization (WMO) version of the IBTrACS global TC database is used for our study, which consists of observations from WMO-designated forecast centers only (Knapp et al. 2010). This global TC best track database is endorsed by the WMO Tropical Cyclone Program. The database is used to obtain the time and position of TC genesis for each TC observed on a global domain. A period of more than 35 years is selected from 1979 (start of the environment dataset, see below) to mid-April 2014 (latest data available at the time of first querying IBTrACS). The IBTrACS database contains about 3200 systems (i.e., low pressure cyclonic systems that have been classified by designated regional and national meteorological agencies as TCs) for this time period, which translates to about 90 systems per year.

The cyclogenesis time is defined as the first time a TC has an observed wind speed greater than or equal to 34 kt (17.5 m s^{-1}). In this research the location of each system before cyclogenesis is important, as we examine the state of the local environment at 12, 24, 36, 48, 60, and 72 h prior to TC genesis. Here, the local environment refers to the surface and atmospheric state of a domain centered on the location of a tropical cloud cluster (TCC), where the central region containing the developing system itself is masked. To extend tracks before the moment of TC genesis the TCC dataset from Hennon et al. (2011) is used. After identification of the cyclogenesis time, the track of the system before TC genesis is obtained from the IBTrACS database. Then, the corresponding TCC is retrieved from the TCC dataset and these track data are used to extend the IBTrACS track. When multiple TCCs are linked to a

specific TC, the TCC with the longest lifetime has been used for track extension. The criteria used for selecting TCCs that develop into TCs are summarized below:

1) If TC wind speed data are unavailable or the wind speed threshold (i.e., 34 kt) is not reached then the corresponding system is not included in this study.
2) For each lead time, a system is only included in this study if the center of its corresponding TCC is located within the geographic boundaries of 30°N–30°S.
3) For each lead time, a system is only included if a location estimate is present in the extended track no more than 6 h away from the requested time. This is to ensure that the TCC has not moved out of the masked region of the retrieved local environment (i.e., the domain centered on the closest available location).
4) The TCs in all basins have been investigated. To capture possible differences in terms of TC genesis in different basins, a BASIN variable has been included in the analysis. Basin classifications used in this study are the western North Pacific, eastern North Pacific, South Pacific, North Atlantic, South Atlantic, north Indian Ocean, and south Indian Ocean.

As a TC developed from each of these TCCs, this selection forms all positive cases in our study. Local environments of the selected TCCs are referred to as the favorable environments. As a result of applying these selection criteria, 2562 systems have been identified from the IBTrACS database for this study (at a 12-h lead time). For longer lead times, the number of valid locations decreases as less TCC track locations can be found within 6 h of the requested date and time. Therefore, the number of selected systems decreases to 2282 at 24-h lead time, 1805 at 36 h, 1423 at 48 h, 1075 at 60 h, and 819 at 72 h.

The same TCC dataset (Hennon et al. 2011) has also been used to select TCCs that do not develop into a TC (i.e., unfavorable environments). This dataset has already been filtered to include only the systems with centers positioned over ocean waters. Thus, the following criteria are used to identify the unfavorable environments:

1) Only systems within the geographic boundaries of 30°N–30°S are included.
2) Only nondeveloping TCCs that occur within the TC season of the respective basin have been selected.
3) One record for each nondeveloping TCC has been randomly selected (instead of all records for a single TCC).

This resulted in 28 264 samples of unique nondeveloping TCCs. Their conditions at the corresponding times and locations are considered as unfavorable for TC genesis. Hence, at 12-h lead time, we compare 30 826 TCCs in total: 28 264 of them are classified as unfavorable environments, while each of the remaining 2562 will develop into a TC within 12 h.

Local environment data for both favorable and unfavorable environments were obtained from the ERA-Interim dataset (Dee et al. 2011). Daily data at 0000, 0600, 1200, and 1800 UTC from 1979 to 2014 were obtained (~400 GB). By subtracting multiples of 12 h from the cyclogenesis time, local environments have been retrieved for when a TC is to be formed within 12 h, between 12 and 24 h, between 24 and 36 h, and so forth until 60–72 h. The selected time at which the local environment is retrieved for each interval is typically near the end of the interval (e.g., in the 12–24-h category the time until cyclogenesis is generally close to 24 h). The ERA-Interim default 0.75° × 0.75° grid was used for the region between 30°N and 30°S. The local environment is then retrieved based on the area centered on the latitude and longitude position of the developing system. Earlier research on tropical cyclogenesis provides indications on the size of the area that could be used. For example, Zeng et al. (2007) and Chen et al. (2011) used a radius of 5° latitude around the storm's center to measure atmospheric variables. McBride and Zehr (1981) concluded that a radius of 6° is optimal to compare vertical wind shear measurements of developing and nondeveloping systems. Furthermore, Peng et al. (2012) and Fu et al. (2012) used area sizes varying from 10° × 10° to 20° × 20° to measure differences between developing and nondeveloping tropical disturbances. Based on these studies, we set the size of the local environment to an approximate 10° × 10° area. To mask representations of incipient TCs in the ERA-Interim dataset, the measurements in a 3.75° × 3.75° area around the center of the retrieved region are discarded. This methodology leads to a comparison of the environments in which TCCs are located, rather than focusing on the developing storm itself. Relative vorticity, potential vorticity, and divergence of the wind field; relative humidity, specific humidity, air temperature, vertical velocity, geopotential and the zonal and meridional components of wind were retrieved. Measurements of these features were obtained at atmospheric pressure levels of 200, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa. The measurement of each feature at each pressure level is referred to as a variable. Other variables were also retrieved from the surface level dataset: SST, mean sea level pressure, 2-m dewpoint temperature, 2-m temperature, low cloud cover, medium cloud cover, high cloud cover, and total cloud cover. Furthermore, the following calculated variables are included. Wind shear is calculated as a vector shear, using both zonal and meridional wind components. The vector shear is calculated

between all combinations of the nine selected atmospheric levels and the result at each combination is included as a variable. The Coriolis parameter is calculated as $2\Omega \sin(\varphi)$ with $\Omega = 7.2921 \times 10^{-5}\,\mathrm{rad\,s^{-1}}$ the rotation rate of Earth and φ the latitude (rad). Absolute vorticity is calculated as the Coriolis parameter plus relative vorticity. Equivalent potential temperature is calculated according to Bolton (1980).

Accessing the ERA-Interim NetCDF files and data processing to retrieve the local environments of TCCs were performed using MATLAB (2015). The resulting dataset, prepared as described above, has been used to compare TCC unfavorable environments where no TC developed with the favorable environments where TC genesis occurred for lead times up to 72 h.

b. *Graphical model structure learning*

The collected local environment data are investigated using a graphical model approach. A graphical model consists of nodes representing random variables and describes the joint probability distribution of all nodes in the graph. This probability distribution can be calculated as the product of certain functions over connected subsets of nodes in the graph. A graphical model also depicts the statistical dependency structure between the nodes, which makes graphical model structure learning an appropriate technique for the discovery of causal relationships. Causal relationships can be obtained by determining the direction of edges after the skeleton of the graph has been learned. Structure learning takes into account not only the individual relationships between variables but also interaction effects between sets of variables. Refer to Koller and Friedman (2009) or Whittaker (1990) for a detailed description of graphical models. Variable selection (also commonly referred to as feature selection) using graphical models has been investigated in various other works (e.g., Yaramakala and Margaritis 2005; Lastra et al. 2011). Furthermore, graphical models have been applied to climate or extreme weather research in several other studies (e.g., Abramson et al. 1996; Ebert-Uphoff and Deng 2012).

In this study a directed acyclic graph (DAG) is constructed. As nodes the variables described in section 2a are used, plus a binary node that encodes whether or not a TC formed; this latter node is referred to as the TC genesis node. For the purpose of exploring the causal structure of variables with the TC genesis node, the entire graph is not required. In theory, the Markov blanket on the TC genesis node should give the complete set of variables that are relevant. However, structure learning is performed for the full graph first, where the remaining structure of the graph (excluding the Markov blanket on the TC genesis node) is used for general validation purposes. The final graph could also be regarded as a temporal–spatial random field. That is, all nodes specify the climate state in spatial terms, except for the TC genesis node, which specifies whether formation will occur at a later time. In addition, the climate state can be updated for each point in time, leading to different probabilistic views of TC genesis for the selected local environments.

Several algorithms have been developed for graphical model structure learning. Investigated in this study are score-based algorithms in combination with a greedy search and constraint-based algorithms, which learn a set of conditional independence requirements and optimize the graph structure using these constraints. After exploration of various algorithms, the constraint-based Peter–Clark (PC) algorithm (Spirtes et al. 2000) was chosen for graphical model structure learning. This algorithm has the theoretical property that it will recover graphs that are faithful to the population distribution given that four assumptions are satisfied. These assumptions are (i) the selected variables form a causally sufficient set, (ii) the same causal relations among the variables hold for every record in the population, (iii) the joint distribution of the observed variables is faithful to a DAG of the causal structure, and (iv) the statistical decisions required by the algorithm are correct for the population (Spirtes et al. 2000, p. 80). It has been noted that often some of the four assumptions cannot be met in practice. The fourth assumption is not met in this study; for example; the consequences and resolutions will be evaluated later in this paper.

The conditional independence tests in graphical model structure learning for continuous variables are generally for testing the linear correlations between the relevant variables, which are assumed to follow a multivariate Gaussian distribution. These distribution and linearity assumptions do not seem appropriate for climate variables that often follow non-Gaussian distributions and additionally, relationships between climatic processes are likely nonlinear. To address the concerns mentioned above nonparametric independence tests are applied to the discretized data. After some empirical exploration each continuous variable is discretized into six bins by default. Independence tests between nodes given a (possibly empty) subset of the remaining nodes have been performed using the G^2 test. A detailed description of this test is provided in appendix A. Using too many bins in the discretization may lead to large type II errors for these independence tests (evidence to be given later in the paper). To assess robustness of the results against the discretization procedure, structure learning has also been performed using five and seven bins. These tests suggest similar rankings of the variables for varying lead times (the results are not presented here).

The combinatorial space of graph structures grows superexponentially as a function of the number of nodes (Robinson 1973). Even though the PC algorithm does not evaluate every DAG, the number of nodes was too large to successfully complete a full run of the algorithm. This was also the result of using independence tests for discrete rather than continuous variables. Therefore, an initial selection was made before graphical model structure learning, based on the strength of the individual relationship between each node and the TC genesis node. Variable evaluation based on mutual information was performed using Weka (Hall et al. 2009). A description of the mutual information criterion and how it has been applied in this study is provided in appendix B. While eliminating variables, care was taken to ensure that the remaining variables give a diverse representation of the climate state, such that the collection of nodes still provides a causally sufficient structure.

Additionally, a challenge for graphical model structure learning is posed by the limited number of TC genesis events. Binning the available data according to level combinations of several nodes often leads to many bins containing no or few genesis observations, implying the conditional independence tests for the involved discretized variables may not be reliably carried out when the conditioning is carried out on several nodes. The resulting false negatives may end in selecting a network structure where edges have been eliminated for variables that are known to have strong influence on TC genesis. This then violates the aforementioned assumption (iv) of the PC algorithm that the statistical decisions required by the algorithm should be correct for the population. A resolution was found to bypass this complication based on the observation that independence testing is successful for conditioning sets that each consist of up to two nodes only. For a conditioning set of two nodes the genesis events are spread over 6^3 levels, which yields on average 11.9 genesis events per level for a 12-h lead time. A conditioning set of three nodes resulted in 6^4 levels or on average less than two positive cases per level (12-h lead time) and even lower for longer lead times, causing precision related complications in some independence tests.

Therefore, the graphical model structure learning approach used in this research is carried out as follows. The PC algorithm is started with the fully connected network. In a first round, edges are considered for removal based on pairwise independence tests for these edges. Then two rounds of conditional independence tests are performed, where conditional independence between each pair of nodes is tested given one and two other nodes, respectively. Afterward, the structure learning algorithm is stopped. Stopping before the full PC algorithm is completed could lead to a network with more edges than the true graph. However, this is not problematic since the aim is to find the nodes having the strongest connection to the TC genesis node for the purpose of variable ranking. Note that for a network of 25 nodes and a fixed lead time a total of about 170 000 independence tests are performed using this approach. Variables are ranked based on the maximum observed p value in all (conditional) independence tests of the corresponding node with the TC genesis node. Using the maximum observed p value is consistent with the methodology of the PC algorithm and emphasizes that it is more important to control the number of false positives instead of false negatives. This maximum observed p value (ζ_i) between variable i and the TC genesis node for a fixed lead time can be formulated as in Eq. (1):

$$\zeta_i = \max_{\forall S \in \Phi_i} \Pr(U_{i|S} > g_{i|S}), \tag{1}$$

with $g_{i|S}$ the value of the G^2 statistic for the independence test between X_i and the TC genesis node given set S; Φ_i is the collection of all sets of zero, one or two nodes considered by the PC algorithm for the X_i and TC genesis nodes. Finally, $U_{i|S}$ is a random variable with a χ^2 distribution and degrees of freedom as defined in appendix A. Note that a lower maximum observed p value indicates a stronger association of the node with TC genesis.

Experiments for graphical model structure learning using varying sets of nodes were performed in R (R Core Team 2014) using a parallel processing adaptation of the PC algorithm by Le et al. (2016). The final structure learning was executed using the PC-stable (Colombo and Maathuis 2014) implementation in the R package *pcalg* (Kalisch et al. 2012).

c. *Validation by logistic regression*

Outcomes of variable ranking based on graphical model structure learning have been validated using a predictive modeling approach. For each considered lead time, all possible combinations of the nodes in the graphical model have been used in a logistic regression model to forecast TC genesis. The validation uses all selected developing systems for a specific lead time together with the 28 264 nondeveloping TCCs. The aim is to find the optimum combination of variables for varying lead times. For use in a logistic regression model some of the variables need to be transformed: the absolute value was taken of relative vorticity, potential vorticity, wind speed, and the Coriolis parameter. Since TCs in both the Northern Hemisphere (NH) and the Southern Hemisphere (SH) are analyzed, these transformations are valid. Predictive performance of TC genesis has been

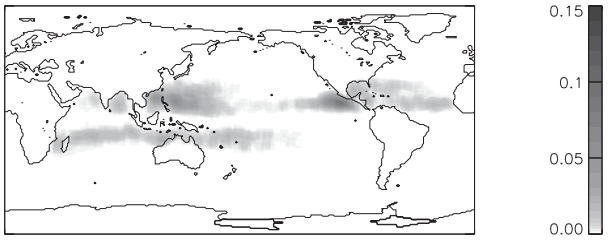

FIG. 1. Average annual TC genesis events per degree latitude squared.

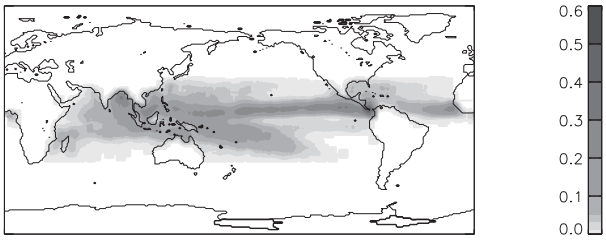

FIG. 2. Average annual nondeveloping TCCs per degree latitude squared.

measured under 10-fold cross validation by the numbers of true negatives, true positives, false negatives, false positives, and the area under the receiver operating characteristic (ROC) curve (Sing et al. 2005). This yields rankings of variable sets that can be used to validate the ranking obtained by the graphical model approach.

3. Results

Figures 1 and 2 show density estimates of the developing and nondeveloping TCCs. Figure 1 displays the annual average number of TC genesis events per degree latitude squared, calculated using all selected developing systems. Figure 2 displays the annual average number of nondeveloping TCCs per degree latitude squared, calculated using all selected nondeveloping TCCs.

Using the retrieved local environment data, the empirical distributions of all selected variables have been analyzed. The atmospheric variables were analyzed at each pressure level as well as by using the average value over all selected pressure levels, which resulted in 150 variables in total. For each variable, one measurement per local environment was obtained by averaging over all corresponding observations in the respective 10° × 10° area, except the observations in the centered 3.75° × 3.75° region. These measurements were used to estimate the distribution for each variable in favorable and unfavorable environments. The goal of this analysis is a visual inspection of whether there exists a difference in distribution between local environments favorable and unfavorable for TC genesis and how this is affected by varying lead time. Figure 3 shows the results of this analysis for a few selected variables, using lead times of 12 and 72 h. The kernel estimate of each probability density function was derived using a Gaussian kernel with bandwidth calculated according to Silverman (1986).

In Fig. 3, the SST distribution shows that TCs form in areas of warm ocean waters and as expected this distribution is relatively stable for varying lead times. Importantly, the distribution of SSTs seems very similar in favorable and

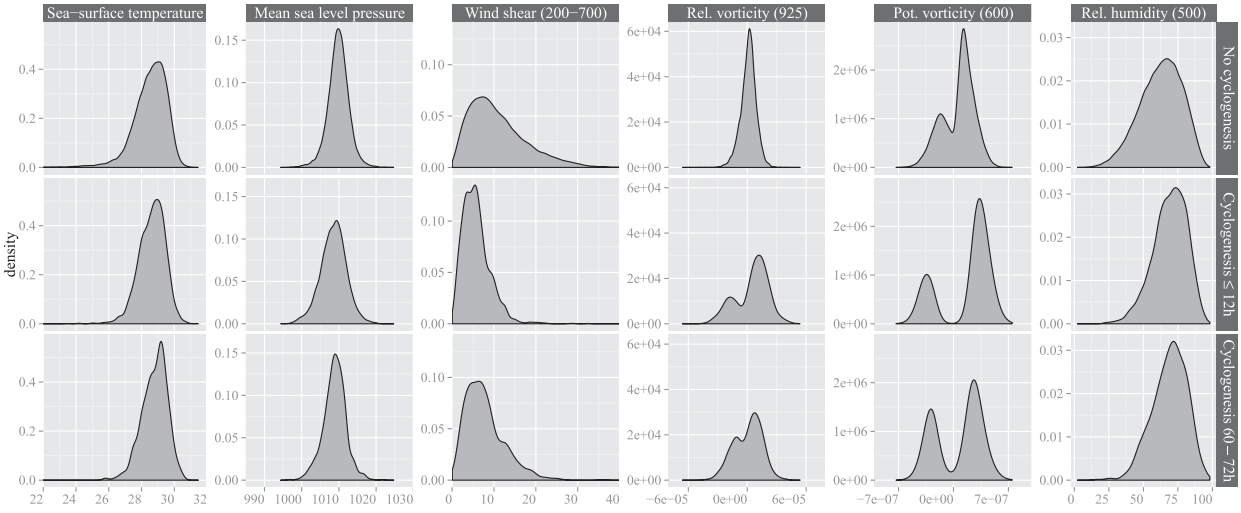

FIG. 3. Kernel estimates of the probability density functions of each column variable in a local environment where no TC forms, where a TC forms within 12 h, and where a TC forms after 60–72 h. The presented variables are SST (°C), mean sea level pressure (hPa), vector wind shear between 200 and 700 hPa ($m\ s^{-1}$), relative vorticity at 925 hPa (s^{-1}), potential vorticity at 600 hPa ($K\ m^2\ kg^{-1}\ s^{-1}$), and relative humidity at 500 hPa (%).

TABLE 1. Mutual information between TC genesis and vector vertical wind shear between different atmospheric pressure levels for a 24-h lead time. Higher mutual information indicates a stronger relationship with TC genesis.

Wind shear (between)	Mutual information
200 and 700 hPa	0.0353
200 and 600 hPa	0.0335
200 and 850 hPa	0.0328
200 and 500 hPa	0.0323
200 and 925 hPa	0.0302
300 and 600 hPa	0.0296
200 and 400 hPa	0.0294
200 and 1000 hPa	0.0293
300 and 700 hPa	0.0293
300 and 500 hPa	0.0292
300 and 400 hPa	0.0270
300 and 850 hPa	0.0262
400 and 500 hPa	0.0249
200 and 300 hPa	0.0244
400 and 600 hPa	0.0232
300 and 925 hPa	0.0225
400 and 700 hPa	0.0212
300 and 1000 hPa	0.0210
400 and 850 hPa	0.0178
500 and 600 hPa	0.0165
500 and 700 hPa	0.0147
400 and 925 hPa	0.0137
400 and 1000 hPa	0.0124
500 and 850 hPa	0.0108
600 and 700 hPa	0.0106
500 and 925 hPa	0.0093
500 and 1000 hPa	0.0086
850 and 1000 hPa	0.0080
850 and 925 hPa	0.0078
600 and 850 hPa	0.0076
700 and 1000 hPa	0.0074
600 and 1000 hPa	0.0074
600 and 925 hPa	0.0072
700 and 925 hPa	0.0063
700 and 850 hPa	0.0054
925 and 1000 hPa	0.0029

unfavorable environments as well. The same can be observed for the mean sea level pressure in both favorable and unfavorable TCC local environments. Vector wind shear tends to be low in favorable environments for TC genesis with its distribution changing slightly over different lead times. However, for longer lead times wind shear in favorable environments still has a clearly different distribution than in environments of nondeveloping TCCs. The two peaks in the distributions of the relative and potential vorticity variables represent TC genesis in the SH (left) and NH (right), with a larger number of TCs forming in the NH. It can be observed that the distributions of relative vorticity at 925 hPa and potential vorticity at 600 hPa slowly revert to the no cyclogenesis distribution for longer lead times, as was also observed for vertical wind shear. Relative humidity at 500 hPa shows clear differences in favorable and unfavorable environments as well, where a higher water content in the atmosphere is indicative of a more favorable environment for TC genesis.

TABLE 2. Mutual information between TC genesis and potential vorticity at different atmospheric pressure levels for a 24-h lead time.

Potential vorticity	Mutual information
600 hPa	0.0754
700 hPa	0.0634
(averaged)	0.0541
500 hPa	0.0533
850 hPa	0.0259
400 hPa	0.0214
300 hPa	0.0091
925 hPa	0.0079
1000 hPa	0.0013
200 hPa	0.0003

Even when considering the variable distributions in the masked central area only (results not presented), the mean sea level pressure distributions in favorable and unfavorable environments are similar. This indicates that the observed differences in relative and potential vorticity distributions are not merely the reflection of a surface low pressure center. In addition, the distributions in Fig. 3 are estimated using measurements farther away from the TCC center and are therefore representative of the environment rather than a description of an incipient cyclone.

a. Variable ranking

After this initial exploratory analysis, attribute evaluation has been performed based on mutual information between each variable and TC genesis. Tables 1, 2, and 3 show the mutual information at a 24-h lead time for the selected atmospheric pressure levels of vector wind shear, potential vorticity, and specific humidity, respectively. Specifically, mutual information for vertical wind shear is calculated using all possible combinations

TABLE 3. Mutual information between TC genesis and specific humidity at different atmospheric pressure levels for a 24-h lead time.

Specific humidity	Mutual information
(averaged)	0.0170
500 hPa	0.0166
400 hPa	0.0147
1000 hPa	0.0147
600 hPa	0.0144
700 hPa	0.0117
300 hPa	0.0114
850 hPa	0.0103
200 hPa	0.0102
925 hPa	0.0089

TABLE 4. Description of nodes in the graphical model.

Node	Description
BASIN	Basin
CORIOLIS	Coriolis parameter
DIV_925	Divergence (measured at 925 hPa)
GEOP_1000	Geopotential (measured at 1000 hPa)
HIGH_CLOUD	Fraction of high cloud cover (pressure levels $\leq 0.45 \times$ surface pressure)
MSLP	Mean sea level pressure
POT_VORT_600	Potential vorticity (measured at 600 hPa)
REL_HUM_500	Relative humidity (measured at 500 hPa)
REL_VORT_925	Relative vorticity (measured at 925 hPa)
SHEAR_200_700	Vector wind shear (measured between 200 and 700 hPa)
SPEC_HUM	Specific humidity (averaged over the selected pressure levels)
SST	Sea surface temperature
TEMP_300	Air temperature (measured at 300 hPa)
TEMP_DEWPOINT	Surface dewpoint temperature
THETA_E	Equivalent potential temperature difference between surface and 500 hPa
VERT_VEL	Vertical velocity (averaged over the selected pressure levels)
WIND_U_200	Zonal component of wind (measured at 200 hPa)

of these pressure levels. These tables show that pressure levels that are close to each other typically have similar results. The analysis has been performed for all lead times from 12 to 72 h and yielded similar results.

An interesting result is that, based on mutual information, specific humidity averaged over the selected pressure levels may be a more useful variable measuring water content in the atmosphere than specific humidity at a particular pressure level. Furthermore, the mutual information level of average specific humidity is also higher than the highest scoring relative humidity variable, which is measured at 500 hPa (mutual information with TC genesis of 0.0125).

The mutual information analyses led to the selection of a set of nodes for graphical model structure learning. As shown in Tables 1–3, sometimes a similar mutual information level was obtained for different pressure levels of the same field. Since the mutual information analysis does not take into account interaction effects with other atmospheric and surface variables, each of these pressure levels was considered. In these cases graphical model structure learning was performed multiple times to identify the best-performing pressure level (results not presented). This variable was then selected to be included as a node in the final graphical model structure learning runs. Table 4 gives an overview of the nodes included in the final runs. It is observed that these nodes cover all six drivers identified by Gray (1975) with slight changes in the selected pressure levels.

Since in this study the aim is to discover the influence of variables on TC genesis, the results focus on the connections of each node with the TC genesis node. Table 5 shows the maximum observed p value of all (conditional) independence tests of each variable with the TC genesis node at various lead times. When the p value is larger than a prespecified significance level α, the corresponding edge is removed from the graph. At that point no more independence tests are performed for this edge; ranking of variables where the edge was eliminated is not considered. Eliminated edges have been highlighted using a bold font style in Table 5.

TABLE 5. Maximum observed p value of all independence tests between variable and TC genesis at varying lead times, as obtained in graphical model structure learning. Variables are sorted by maximum p value at a 12-h lead time, where lower numbers indicate a stronger connection with TC genesis. The boldface font indicates the edge of the variable to the TC genesis node was removed based on $\alpha = 0.01$.

Node	12 h	24 h	36 h	48 h	60 h	72 h
POT_VORT_600	2.2×10^{-191}	5.7×10^{-162}	4.6×10^{-125}	3.4×10^{-72}	3.4×10^{-35}	3.7×10^{-15}
SHEAR_200_700	2.3×10^{-158}	4.3×10^{-112}	2.5×10^{-61}	2.5×10^{-24}	2.3×10^{-7}	1.5×10^{-5}
REL_VORT_925	1.9×10^{-115}	1.4×10^{-115}	4.0×10^{-80}	1.5×10^{-44}	1.6×10^{-23}	3.1×10^{-11}
DIV_925	4.4×10^{-7}	7.1×10^{-8}	$\mathbf{1.6 \times 10^{-2}}$	$\mathbf{7.8 \times 10^{-2}}$	$\mathbf{1.5 \times 10^{-1}}$	$\mathbf{5.7 \times 10^{-2}}$
TEMP_300	8.7×10^{-5}	$\mathbf{5.3 \times 10^{-2}}$	$\mathbf{2.6 \times 10^{-2}}$	$\mathbf{3.2 \times 10^{-1}}$	$\mathbf{1.4 \times 10^{-1}}$	$\mathbf{1.9 \times 10^{-1}}$
WIND_U_200	2.4×10^{-4}	2.2×10^{-4}	4.0×10^{-3}	$\mathbf{5.4 \times 10^{-2}}$	5.6×10^{-3}	5.9×10^{-4}
BASIN	5.2×10^{-4}	1.1×10^{-9}	6.7×10^{-11}	8.9×10^{-13}	2.8×10^{-5}	8.5×10^{-6}
VERT_VEL	5.5×10^{-3}	3.0×10^{-5}	$\mathbf{5.3 \times 10^{-2}}$	$\mathbf{5.4 \times 10^{-2}}$	$\mathbf{1.0 \times 10^{-1}}$	$\mathbf{3.7 \times 10^{-1}}$
SST	$\mathbf{1.4 \times 10^{-2}}$	$\mathbf{2.6 \times 10^{-2}}$	8.3×10^{-3}	$\mathbf{2.7 \times 10^{-2}}$	$\mathbf{1.5 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{-1}}$
REL_HUM_500	$\mathbf{1.8 \times 10^{-2}}$	$\mathbf{1.3 \times 10^{-1}}$	$\mathbf{1.5 \times 10^{-2}}$	$\mathbf{4.3 \times 10^{-1}}$	$\mathbf{4.3 \times 10^{-1}}$	$\mathbf{1.6 \times 10^{-1}}$
HIGH_CLOUD	$\mathbf{1.8 \times 10^{-2}}$	$\mathbf{6.5 \times 10^{-2}}$	$\mathbf{2.3 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{-2}}$	$\mathbf{3.8 \times 10^{-1}}$	$\mathbf{1.3 \times 10^{-1}}$
TEMP_DEWPOINT	$\mathbf{2.4 \times 10^{-2}}$	2.8×10^{-4}	$\mathbf{1.1 \times 10^{-2}}$	$\mathbf{6.4 \times 10^{-2}}$	$\mathbf{4.1 \times 10^{-1}}$	$\mathbf{1.7 \times 10^{-2}}$
THETA_E	$\mathbf{8.4 \times 10^{-2}}$	$\mathbf{3.4 \times 10^{-1}}$	$\mathbf{3.1 \times 10^{-2}}$	$\mathbf{7.1 \times 10^{-2}}$	$\mathbf{5.2 \times 10^{-2}}$	$\mathbf{2.9 \times 10^{-1}}$
SPEC_HUM	$\mathbf{1.1 \times 10^{-1}}$	$\mathbf{2.8 \times 10^{-1}}$	$\mathbf{5.5 \times 10^{-1}}$	$\mathbf{4.2 \times 10^{-2}}$	$\mathbf{1.7 \times 10^{-1}}$	$\mathbf{1.1 \times 10^{-1}}$
CORIOLIS	$\mathbf{5.1 \times 10^{-1}}$	$\mathbf{2.0 \times 10^{-2}}$	$\mathbf{3.0 \times 10^{-2}}$	$\mathbf{9.0 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{1.0 \times 10^{0}}$
GEOP_1000	$\mathbf{9.7 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{9.7 \times 10^{-1}}$	$\mathbf{5.9 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{1.1 \times 10^{-1}}$
MSLP	$\mathbf{9.8 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{9.0 \times 10^{-1}}$	$\mathbf{4.7 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{7.9 \times 10^{-2}}$

TABLE 6. Summary of highest-ranked variables per lead time, as obtained through graphical model structure learning. Ranks are only displayed when the edge of a variable with the TC genesis node has not been removed.

Rank	12 h	24 h	36 h	48 h	60 h	72 h
1	POT_VORT_600	POT_VORT_600	POT_VORT_600	POT_VORT_600	POT_VORT_600	POT_VORT_600
2	SHEAR_200_700	REL_VORT_925	REL_VORT_925	REL_VORT_925	REL_VORT_925	REL_VORT_925
3	REL_VORT_925	SHEAR_200_700	SHEAR_200_700	SHEAR_200_700	SHEAR_200_700	BASIN
4	DIV_925	BASIN	BASIN	BASIN	BASIN	SHEAR_200_700
5	TEMP_300	DIV_925	WIND_U_200	—	WIND_U_200	WIND_U_200

It is evident that the maximum observed p value mostly increases with lead time, indicating that forecasting is more difficult at longer lead times. This also leads to a decrease in the number of edges with the TC genesis node for longer lead times, indicating that some variables can only be used for short-term forecasting.

A summary of the most highly ranked variables at various lead times is presented in Table 6. To show the change in relative importance of drivers for TC genesis over time, rankings of the variables are also plotted in Fig. 4. Both Table 6 and Fig. 4 show that POT_VORT_600, REL_VORT_925, and SHEAR_200_700 have a strong connection with TC genesis at all lead times. This is also the case for the BASIN and WIND_U_200 variables. At a 48-h lead time the edge of WIND_U_200 with the TC genesis node was removed based on a conditional independence test involving SHEAR_200_700. This indicates that SHEAR_200_700 already incorporates some of the same signals encapsulated in WIND_U_200 with respect to TC genesis. Finally, DIV_925, TEMP_300, and VERT_VEL are important variables at shorter lead times.

b. Validation

To validate the above findings predictive modeling using logistic regression was performed, where performance was measured in cross validation for a large number of models (i.e., using all possible combinations of the selected variables). Naturally, the best performing set of variables in such a predictive model is expected to include the top-ranked variables. Table 7 shows the best-performing models per lead time using four variables, based on their values of the area under the ROC curve (AUC).

From Table 7 it can be observed that POT_VORT_600 and SHEAR_200_700 are selected for all lead times. REL_VORT_925 is in the best-performing model at 12-, 24-, and 36-h lead times, while BASIN and SPEC_HUM are part of the best-performing model at longer lead times only. In addition, Fig. 5 shows the performance results of the best-performing logistic regression model for forecasting TC genesis at varying lead times. Plots on the left are ROC curves; the corresponding AUC values are presented in Table 8. The ROC curve of a perfect model (i.e., AUC = 1) passes through the point (0, 1),

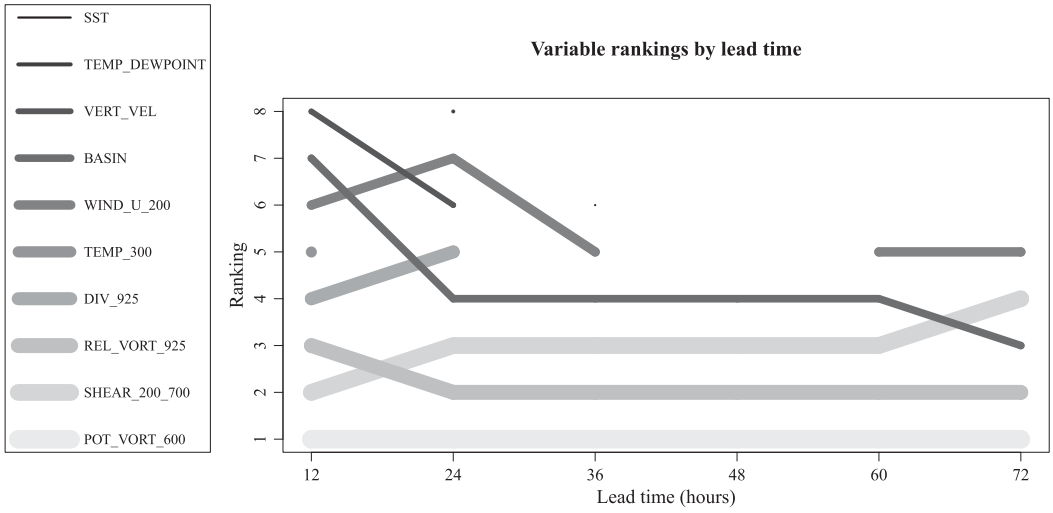

FIG. 4. Progression of variable rankings for changing lead times, as obtained through graphical model structure learning. A lower rank indicates a stronger connection with TC genesis. Ranks are only displayed when the edge of a variable with the TC genesis node has not been removed.

TABLE 7. Variables selected in the top logistic regression models with four variables. The logistic regression models are used for model validation.

Lead time	Selected variables in top logistic regression model (random order)
≤12 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, REL_VORT_925
12–24 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, REL_VORT_925
24–36 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, REL_VORT_925
36–48 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, BASIN
48–60 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, SPEC_HUM
60–72 h	POT_VORT_600, SHEAR_200_700, CORIOLIS, BASIN

while a random forecast would lead to performance close to the dotted line. The larger the area under the ROC curve, the better the forecasting performance. The right column in Fig. 5 contains density histograms of TC genesis model predictions for both favorable and unfavorable environments. These plots show how well the logistic regression model splits favorable and unfavorable environments for TC genesis.

At all lead times CORIOLIS is part of the best-performing logistic regression model. In contrast, this variable is consistently removed in the graphical model structure learning process in conditional independence tests involving POT_VORT_600 for all lead times. This indicates that the CORIOLIS signal is already incorporated in POT_VORT_600. The three-variable model at a 12-h lead time using POT_VORT_600, SHEAR_200_700, and REL_VORT_925 scores an AUC of 0.9422 in cross validation, with the number of true positives equal to 1114. Hence, the performance is very similar to the best four-variable logistic regression model that also incorporates CORIOLIS (AUC of 0.9458 and 1192 true positives). In addition, the selection of CORIOLIS in the top logistic regression models could be influenced by our inclusion criteria for nondeveloping TCCs.

More generally, when considering the performance of the top four variables determined during graphical model structure learning in a logistic regression model, the results are positive. These models show very similar performance as the top logistic regression model at the corresponding lead time. For example, at a 24-h lead time the logistic regression model using POT_VORT_600, REL_VORT_925, SHEAR_200_700, and BASIN ranks third of all four-variable models. The model has an AUC of 0.9305, compared to 0.9313 of the top logistic regression model. Therefore, the validation results are similar to the variable rankings that were obtained from the graphical

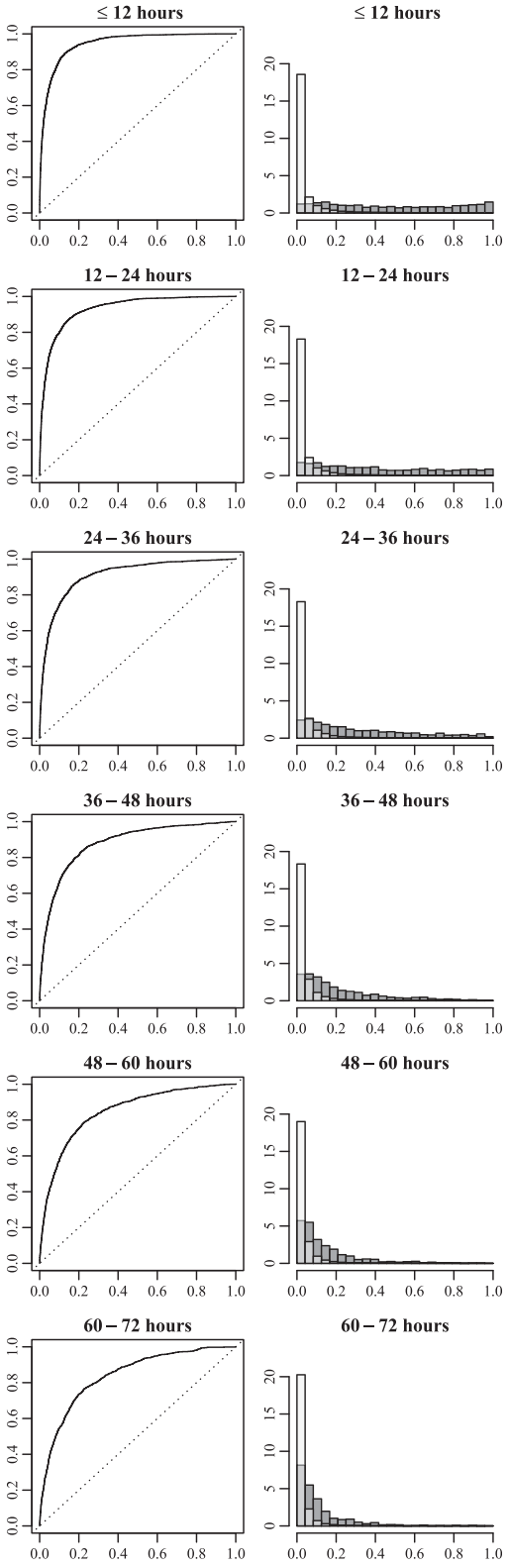

FIG. 5. (left) ROC curves with the false positive rate on the *x* axis and true positive rate on the *y* axis. (right) Two density histograms of model predictions per plot with white for unfavorable environments and gray for favorable environments for TC genesis. Both histograms are displayed in one plot, leading to areas where the density histograms overlap.

TABLE 8. The number of true negatives (TN), false positives (FP), false negatives (FN), true positives (TP), and the AUC statistic for the top logistic regression model with four variables at a specific lead time. TN, FP, FN, and TP correspond to a cutoff value of 0.5 in the histograms in Fig. 5. This table shows performance in 10-fold cross validation for varying lead times. The logistic regression models are used for model validation purposes.

Lead time	TN	FP	FN	TP	AUC
≤12 h	27 788	476	1370	1192	0.9458
12–24 h	27 858	406	1399	883	0.9313
24–36 h	27 995	269	1349	456	0.9102
36–48 h	28 121	143	1252	171	0.8791
48–60 h	28 198	66	1025	50	0.8479
60–72 h	28 239	25	804	15	0.8379

model structure learning presented in this study (see Table 6 and Fig. 4). Note that the variable ranking results based on graphical model structure learning are considered more important, as the logistic regression models do not fully consider interaction effects between variables and nonlinear relationships with TC genesis.

The AUC values in Table 8 are high, indicating a high probability that a randomly drawn local environment containing a developing TCC will get a higher prediction score than a randomly drawn local environment containing a nondeveloping TCC. Out of 30 826 TCC environments, 1192 TCCs were correctly identified as environments where a TC would form within 12 h. This is equivalent to 47% of the genesis events that were correctly forecasted at this lead time, combined with a false positive rate of 1.68%, using only local environment information and excluding measurements in the TCC's central region. For longer lead times lower accuracy is obtained, but environments where a TC formed are generally still assigned a higher genesis probability than the unfavorable environments (see also the density histograms in Fig. 5).

For completeness, Eq. (2) lists the fitted logistic regression model obtained during the validation exercise for a 24-h lead time. Note that the purpose for this analysis was not to identify the optimal forecasting model, but to identify which variables are important for constructing a predictive model. For example, higher forecasting performance might be obtained by considering interaction effects or different modeling techniques. Also note that higher performance using different coefficients can be obtained without masking of the local environment center (results not presented). For Eq. (2) the model is calibrated on the full training set for cyclogenesis in 12–24 h; this equation could be used to give an indication of the TC genesis probability at this lead time. The variables used in Eq. (2) are the absolute value of POT_VORT_600 ($\mathrm{K\,m^2\,kg^{-1}\,s^{-1}}$), the absolute value of REL_VORT_925 ($\mathrm{s^{-1}}$), SHEAR_200_700 ($\mathrm{m\,s^{-1}}$), and the absolute value of CORIOLIS ($\mathrm{rad\,s^{-1}}$). Variable measurements are obtained using the average value over a 10° × 10° area with a centered 3.75° × 3.75° mask (see section 2a), except for CORIOLIS, which is calculated based on the location of the TCC center:

$$\Pr(\text{TC genesis within 12–24 h}) = \frac{1}{1+e^{-z}} \text{ with}$$
$$z = -4.259\,812 + 20\,034\,940|\text{POT\_VORT\_600}| + 73\,645.39|\text{REL\_VORT\_925}| - 0.240\,207\,3\text{SHEAR\_200\_700} - 74\,129.84|\text{CORIOLIS}|. \quad (2)$$

4. Discussion

This study performed variable selection for TC genesis modeling with a broad selection of variables from ERA-Interim data as the starting point. Using the approach described in section 2 the initial selection of 150 variables is brought down to a selection of top variables. The top-ranked variables include a subset of the six drivers suggested by Gray (1975), which provides support for the performed procedures. In addition, some other variables also performed very well. In particular, potential vorticity at 600 hPa ranks high as variable at all investigated lead times.

Interestingly, the edge between SST and TC genesis is eliminated at most lead times. At these lead times, the conditional independence test involving the SPEC_HUM node fails to reject the null hypothesis of independence between SST and TC genesis, leading to the removal of the edge between SST and TC genesis (see Table 5). This indicates that SST does not provide valuable extra information related to TC genesis given this other node. Hence, the graphical model implies that average specific humidity is a more useful indicator than SST at these lead times (although the edge between SPEC_HUM and TC genesis is also removed later in the process). In the mutual information analysis, where we purely consider the individual relationship of these variables with TC genesis (without interaction effects), SPEC_HUM also consistently ranks higher than SST.

Therefore, we postulate that SST could be important with respect to the formation of a TCC and with respect to the intensity that a TC can reach once formed; however, whether TC genesis will occur from a TCC within the next three days is not largely impacted by the SST conditions at that moment.

Furthermore, based on the mutual information analyses, specific humidity averaged over the atmospheric pressure levels (or equivalently, total water vapor content in the atmosphere) shows to be a more useful variable than relative humidity in the middle troposphere. However, the edges of SPEC_HUM and REL_HUM_500 with the TC genesis node are eventually also removed from the graphical model.

Some previous studies (e.g., Emanuel and Nolan 2004; Chand and Walsh 2011) use vector wind shear between 200 and 850 hPa for TC genesis forecasting. Both the analysis based on mutual information (see Table 1) and a comparison of graphical model structure learning runs using SHEAR_200_700 and SHEAR_200_850 (results not presented) indicate SHEAR_200_700 is a more powerful variable to separate developing from nondeveloping TCCs. The difference between both variables is not very large, but this global analysis indicates that SHEAR_200_700 is preferred over SHEAR_200_850.

The results of this research show that vorticity (POT_VORT_600 and REL_VORT_925) in the local environment surrounding a TCC is a key contributor to TC genesis. A potential limitation of this research could be that some vortex representation of an incipient cyclone is present in the ERA-Interim reanalysis dataset, even when masking the center of the tropical disturbance. However, other research also indicates that it is more likely that the origin of large levels of vorticity (favorable for TC genesis) at the TCC development stage are external to the system, which is in line with our results (e.g., McBride and Zehr 1981; Sippel et al. 2006).

Variable selection in relation to TC genesis is also performed by Peng et al. (2012) and Fu et al. (2012). These studies focus on specific basins and use a different methodology than the current study, which is referred to as the box difference index (BDI). BDI is based on the difference in mean value of a variable in favorable and unfavorable environments for TC genesis, corrected for the variable's variability in these environments. Comparing conclusions of these studies to the current study, Fu et al. (2012) find that a key driver for forecasts in the western North Pacific is low-level relative vorticity. Peng et al. (2012) select the water vapor content over multiple pressure levels instead of relative or specific humidity at a particular pressure level. Using a global perspective, a different dataset and different mathematical techniques, our study confirms these findings.

Differences include the relatively high importance of SST in the North Atlantic basin (Peng et al. 2012) and relatively low importance of wind shear (based on only the zonal component) in the western North Pacific (Fu et al. 2012). In our research, both the selection approach using graphical model structure learning and the validation using predictive modeling indicate that wind shear is a key factor. Note that wind shear is based here on wind vectors, using both zonal and meridional components. In our opinion a limitation of the BDI approach is that it only considers the performance of variables in isolation. Whether or not some variables may possess a similar signal is hence not taken into account in the variable ranking. In our current study however, this is taken into account using conditional independence tests.

Basin-specific variable selection has not been investigated in this study and would require further research. The high ranking of BASIN at all lead times does show that differences between basins exist. Ideally the diversity of the physical variables between basins should address this, but this would require further research. In addition, the analysis for variable selection with a 72-h lead time could form the basis for an investigation of predictive modeling on longer time scales. Such an analysis is likely to require the consideration of additional variables.

Finally, the structure learning performed in this research focused on retrieving the full graph, which allowed for a more detailed validation of the results. However, future applications of this methodology could focus on learning just the local graphical model connected to the TC genesis node (i.e., the Markov blanket). This approach may also eliminate the need for attribute selection based on mutual information.

5. Summary

In this study the potential of various variables at different atmospheric pressure levels was explored for short-term forecasting of TC genesis. The IBTrACS database was used to identify the genesis time and position of over 2500 TCs on a global domain. Tracks before cyclogenesis were extended using a TCC dataset. In addition, over 28 000 nondeveloping TCCs were selected. Using ERA-Interim data the state of each local environment around a TCC was obtained, while the central region of the tropical disturbance itself was masked. Environments containing developing and nondeveloping TCCs were compared to identify characteristics of favorable and unfavorable environments for TC genesis.

This study adopted a structured approach to obtain rankings of variables that describe these local environments for lead times from less than 12 to 72 h. First, the distributions of variables in favorable and unfavorable

environments for TC genesis were explored using kernel estimates of the probability density functions. Then, variables were ranked based on mutual information for an initial selection of potentially favorable variables. At this stage the individual relationship of a variable with TC genesis was measured to eliminate variables/pressure levels with low performance. This initial selection formed a set of nodes for graphical model structure learning using the PC algorithm. Structure learning identified the variables that have the strongest influence on TC genesis, while taking into account the interrelationship with other variables. Variables were then ranked based on the maximum observed p value among all (conditional) independence tests of the variable with the TC genesis node, where a lower maximum observed p value indicated a stronger influence on TC genesis. Finally, the findings were validated using predictive modeling with logistic regression. Performance of all combinations of variables used in the graphical model has been measured using the AUC statistic in cross validation for various lead times. The variables that were used in the best-performing models are in line with the variable rankings obtained by graphical model structure learning. Furthermore, the genesis forecasting models have shown good out-of-sample performance, especially for short lead times.

The results indicate that potential vorticity at 600 hPa, relative vorticity at 925 hPa, and (vector) vertical wind shear between 200 and 700 hPa are the most highly ranked variables. The analyses show that measurements at these pressure levels are the most powerful to separate developing and nondeveloping TCCs. The BASIN variable shows a relatively high ranking at all lead times, which indicates that differences between basins exist. Additionally, the zonal wind speed at 200 hPa can be useful to consider, although some of the signals in this variable are already captured by the vector vertical wind shear between 200 and 700 hPa. We postulate that SST can be influential on the formation of a TCC and the intensity of a TC (once formed), but TC genesis within a 3-day time frame mainly depends on other variables. Some patterns in ranking changes have been observed for varying lead times. For example, in addition to the variables mentioned above, divergence at 925 hPa, air temperature at 300 hPa, and average vertical velocity perform well at short lead times. As a result, the set of variables for which the edges with the TC genesis node have not been removed, shrinks from eight variables at a 12-h lead time to five variables at a 72-h lead time.

The results are consistent with the main outcomes of other earlier studies on variable ranking. Furthermore, the validation performed in our study shows that the variable ranking can be used successfully for predictive modeling of TC genesis at various lead times. Hence, in the future this (methodology for) variable ranking can be used as a building block for creating predictive indices or models.

Acknowledgments. The authors thank Dr. Andrew J. Dowdy for his contributions to this research. The authors would also like to acknowledge the valuable feedback of the editor and three anonymous reviewers, which helped us to improve the quality of the original manuscript.

APPENDIX A

Independence Testing

This appendix describes the (conditional) independence tests for discrete variables that are performed in graphical model structure learning using the PC algorithm. Independence tests are based on the G^2 test statistic (see Neapolitan 2004).

Define X_i for $i = 1, \ldots, n$ as a set of nodes with n being the total number of nodes in a graphical model. Let M be the number of observations in the data used to learn the structure of the graphical model. Also let T_i^a be a random variable denoting the number of times $X_i = a$, with τ_i^a the observation of T_i^a in the data, and let T_{ij}^{ab} be the number of times $X_i = a$ and $X_j = b$ with τ_{ij}^{ab} similarly defined, and so forth.

The G^2 test is for testing independence between nodes X_i and X_j based on comparing the observed and expected frequencies. The G^2 statistics for pairwise independence testing, independence testing given one node (X_k), and testing given two nodes (X_k, X_l) are given in Eqs. (A1), (A2), and (A3), respectively:

$$G^2 = 2\sum_{a,b} \tau_{ij}^{ab} \ln\left(\frac{\tau_{ij}^{ab} M}{\tau_i^a \tau_j^b}\right), \tag{A1}$$

where the summation is over all pairs of (a, b),

$$G^2 = 2\sum_{c}\sum_{a,b} \tau_{ijk}^{abc} \ln\left(\frac{\tau_{ijk}^{abc} \tau_k^c}{\tau_{ik}^{ac} \tau_{jk}^{bc}}\right), \tag{A2}$$

stratified sum per $c \in \mathscr{X}_k$ with $\mathscr{X}_k$ the collection of levels of X_k,

$$G^2 = 2\sum_{c,d}\sum_{a,b} \tau_{ijkl}^{abcd} \ln\left(\frac{\tau_{ijkl}^{abcd} \tau_{kl}^{cd}}{\tau_{ikl}^{acd} \tau_{jkl}^{bcd}}\right), \tag{A3}$$

stratified sum per $c \in \mathscr{X}_k$ and $d \in \mathscr{X}_l$.

Under the null hypothesis, each G^2 statistic in Eqs. (A1)–(A3) follows a χ^2 distribution with respective degrees of freedom f as follows:

$$f = (r_i - 1)(r_j - 1), \quad \text{or} \tag{A4}$$

$$f = (r_i - 1)(r_j - 1)r_k, \quad \text{or} \tag{A5}$$

$$f = (r_i - 1)(r_j - 1)r_k r_l, \tag{A6}$$

with r_i being the number of levels of X_i,

and so forth.

Suppose g is the value of G^2 computed based on the data. Then the null hypothesis of (conditional) independence is rejected if $\Pr(G^2 > g) < \alpha$ for a prespecified significance level α.

APPENDIX B

Mutual Information Criterion

This appendix provides a short description of the mutual information criterion (e.g., Cover and Thomas 1991) and how it is applied to evaluate variables for TC genesis forecasting. The mutual information $I(X;Y)$ between two discrete random variables X and Y can be described as a function of entropies [see Eq. (B1)], where entropy is a measure of uncertainty of a variable:

$$I(X;Y) = H(X) - H(X \mid Y), \text{with} \tag{B1}$$

$$H(X) = -\sum_{x \in \mathscr{X}} p(x) \log\, p(x) \text{ the entropy of } X,$$

$p(x)$ is the probability mass function (pmf),

$\mathscr{X}$ is the collection of distinct values of X,

and the logarithm has base 2, and

$$H(X \mid Y) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x,y) \log\, p(x \mid y) \text{ with}$$

$p(x,y)$ the joint pmf of (X,Y), and

$p(x \mid y)$ the conditional pmf of X given Y.

As mutual information is symmetric, this can also be written as $I(X;Y) = H(Y) - H(Y \mid \mathrm{X})$. In this study, the mutual information between each variable X_i and the TC genesis node (i.e., G) is calculated as in Eq. (B2):

$$I(X_i;G) = H(G) - H(G \mid \mathrm{X}_i) \quad \text{for} \quad i = 1, \ldots, n \tag{B2}$$

with n the number of variables that are investigated.

This can be regarded as the reduction in uncertainty of TC genesis due to knowledge of X_i. A ranking is obtained by sorting $\{[X_i, I(X_i;G)]: i \in \{1, \ldots, n\}\}$ on $I(X_i;G)$, where a higher mutual information indicates a stronger association. Based on this ranking, it is observed which pressure levels for a specific field (e.g., relative humidity) have the highest potential for forecasting TC genesis and should be included as a node in graphical model structure learning.

REFERENCES

Abramson, B., J. Brown, W. Edwards, A. Murphy, and R. L. Winkler, 1996: Hailfinder: A Bayesian system for forecasting severe weather. *Int. J. Forecasting*, **12**, 57–71, doi:10.1016/0169-2070(95)00664-8.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053, doi:10.1175/1520-0493(1980)108<1046:TCOEPT>2.0.CO;2.

Camargo, S. J., K. A. Emanuel, and A. H. Sobel, 2007: Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis. *J. Climate*, **20**, 4819–4834, doi:10.1175/JCLI4282.1.

Chand, S. S., and K. J. E. Walsh, 2011: Forecasting tropical cyclone formation in the Fiji region: A probit regression approach using Bayesian fitting. *Wea. Forecasting*, **26**, 150–165, doi:10.1175/2010WAF2222452.1.

Charney, J. G., and A. Eliassen, 1964: On the growth of the hurricane depression. *J. Atmos. Sci.*, **21**, 68–75, doi:10.1175/1520-0469(1964)021<0068:OTGOTH>2.0.CO;2.

Chen, P., H. Yu, and J. C. L. Chan, 2011: A western North Pacific tropical cyclone intensity prediction scheme. *Acta Meteor. Sin.*, **25**, 611–624, doi:10.1007/s13351-011-0506-9.

Colombo, D., and M. H. Maathuis, 2014: Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, **15** (1), 3921–3962.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. John Wiley & Sons, 542 pp.

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, doi:10.1002/qj.828.

DeMaria, M., J. A. Knaff, and B. H. Connell, 2001: A tropical cyclone genesis parameter for the tropical Atlantic. *Wea. Forecasting*, **16**, 219–233, doi:10.1175/1520-0434(2001)016<0219:ATCGPF>2.0.CO;2.

Ebert-Uphoff, I., and Y. Deng, 2012: A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer. *Geophys. Res. Lett.*, **39**, L19701, doi:10.1029/2012GL053269.

Emanuel, K. A., 1987: An air-sea interaction model of intraseasonal oscillations in the tropics. *J. Atmos. Sci.*, **44**, 2324–2340, doi:10.1175/1520-0469(1987)044<2324:AASIMO>2.0.CO;2.

——, 2013: Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century. *Proc. Natl. Acad. Sci. USA*, **110**, 12 219–12 224, doi:10.1073/pnas.1301293110.

——, and D. S. Nolan, 2004: Tropical cyclone activity and the global climate system. *26th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 10A.2. [Available online at http://ams.confex.com/ams/26HURR/techprogram/paper_75463.htm.]

Fu, B., M. S. Peng, T. Li, and D. E. Stevens, 2012: Developing versus nondeveloping disturbances for tropical cyclone formation. Part II: Western North Pacific. *Mon. Wea. Rev.*, **140**, 1067–1080, doi:10.1175/2011MWR3618.1.

Gray, W. M., 1975: Tropical cyclone genesis. Atmospheric Science Paper 234, Dept. of Atmospheric Science, Colorado State University, 121 pp. [Available online at http://hdl.handle.net/10217/247.]

——, 1998: The formation of tropical cyclones. *Meteor. Atmos. Phys.*, **67**, 37–69, doi:10.1007/BF01277501.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, 2009: The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.*, **11**, 10–18, doi:10.1145/1656274.1656278.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting*, **28**, 1423–1445, doi:10.1175/WAF-D-13-00008.1.

Hennon, C. C., C. N. Helms, K. R. Knapp, and A. R. Bowen, 2011: An objective algorithm for detecting and tracking tropical cloud clusters: Implications for tropical cyclogenesis prediction. *J. Atmos. Oceanic Technol.*, **28**, 1007–1018, doi:10.1175/2010JTECHA1522.1.

Kalisch, M., M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, 2012: Causal inference using graphical models with the R package pcalg. *J. Stat. Software*, **47**, 1–26, doi:10.18637/jss.v047.i11.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, doi:10.1175/2009BAMS2755.1.

Koller, D., and N. Friedman, 2009: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 1280 pp.

Lastra, G., O. Luaces, J. R. Quevedo, and A. Bahamonde, 2011: Graphical feature selection for multilabel classification tasks. *Advances in Intelligent Data Analysis X*, J. Gama, E. Bradley, and J. Hollmén, Eds., Springer, 246–257, doi:10.1007/978-3-642-24800-9_24.

Le, T. D., T. Hoang, J. Li, L. Liu, H. Liu, and S. Hu, 2016: A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi:10.1109/TCBB.2016.2591526, in press.

MATLAB, 2015: Release 2015a. The MathWorks, Inc., Natick, MA.

McBride, J. L., and R. Zehr, 1981: Observational analysis of tropical cyclone formation. Part II: Comparison of non-developing versus developing systems. *J. Atmos. Sci.*, **38**, 1132–1151, doi:10.1175/1520-0469(1981)038<1132:OAOTCF>2.0.CO;2.

Murakami, H., and Coauthors, 2012: Future changes in tropical cyclone activity projected by the new high-resolution MRI-AGCM. *J. Climate*, **25**, 3237–3260, doi:10.1175/JCLI-D-11-00415.1.

Neapolitan, R. E., 2004: *Learning Bayesian Networks*. Pearson, 674 pp.

Peng, M. S., B. Fu, T. Li, and D. E. Stevens, 2012: Developing versus nondeveloping disturbances for tropical cyclone formation. Part I: North Atlantic. *Mon. Wea. Rev.*, **140**, 1047–1066, doi:10.1175/2011MWR3617.1.

R Core Team, 2014: R: A language and environment for statistical computing, version 3.2.0. R Foundation for Statistical Computing. [Available online at http://www.R-project.org/.]

Robinson, R. W., 1973: Counting labeled acyclic digraphs. *New Directions in the Theory of Graphs*, F. Harary, Ed., Academic Press, 239–273.

Schreck, C. J., J. Molinari, and A. Aiyyer, 2012: A global view of equatorial waves and tropical cyclogenesis. *Mon. Wea. Rev.*, **140**, 774–788, doi:10.1175/MWR-D-11-00110.1.

Shen, B.-W., W.-K. Tao, W. K. Lau, and R. Atlas, 2010: Predicting tropical cyclogenesis with a global mesoscale model: Hierarchical multiscale interactions during the formation of tropical cyclone Nargis (2008). *J. Geophys. Res.*, **115**, D14102, doi:10.1029/2009JD013140.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 176 pp.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2005: ROCR: Visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941, doi:10.1093/bioinformatics/bti623.

Sippel, J. A., J. W. Nielsen-Gammon, and S. E. Allen, 2006: The multiple-vortex nature of tropical cyclogenesis. *Mon. Wea. Rev.*, **134**, 1796–1814, doi:10.1175/MWR3165.1.

Slade, S. A., and E. D. Maloney, 2013: An intraseasonal prediction model of Atlantic and east Pacific tropical cyclone genesis. *Mon. Wea. Rev.*, **141**, 1925–1942, doi:10.1175/MWR-D-12-00268.1.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.*, **3**, 846–849, doi:10.1038/ngeo1004.

Spirtes, P., C. N. Glymour, and R. Scheines, 2000: *Causation, Prediction, and Search*. 2nd ed. MIT Press, 543 pp.

Vecchi, G. A., and Coauthors, 2013: Multiyear predictions of North Atlantic hurricane frequency: Promise and limitations. *J. Climate*, **26**, 5337–5357, doi:10.1175/JCLI-D-12-00464.1.

——, and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, doi:10.1175/JCLI-D-14-00158.1.

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, doi:10.1175/2010MWR3343.1.

Ward, G. F. A., 1995: Prediction of tropical cyclone formation in terms of sea-surface temperature, vorticity and vertical wind shear. *Aust. Meteor. Mag.*, **44**, 61–70.

Whittaker, J., 1990: *Graphical Models in Applied Multivariate Statistics*. Wiley, 466 pp.

Wijnands, J. S., G. Qian, K. L. Shelton, R. J. B. Fawcett, J. C. L. Chan, and Y. Kuleshov, 2015: Seasonal forecasting of tropical cyclone activity in the Australian and the South Pacific Ocean regions. *Math. Climate Wea. Forecasting*, **1** (1), 21–42, doi:10.1515/mcwf-2015-0002.

Yaramakala, S., and D. Margaritis, 2005: Speculative Markov blanket discovery for optimal feature selection. *Proc. Fifth IEEE Int. Conf. on Data Mining*, Houston, TX, IEEE, doi:10.1109/ICDM.2005.134.

Zeng, Z., Y. Wang, and C.-C. Wu, 2007: Environmental dynamical control of tropical cyclone intensity—An observational study. *Mon. Wea. Rev.*, **135**, 38–59, doi:10.1175/MWR3278.1.

Zhang, W., B. Fu, M. S. Peng, and T. Li, 2015: Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. *Wea. Forecasting*, **30**, 446–454, doi:10.1175/WAF-D-14-00023.1.