Topographic Enhancement of Tropical Cyclone Precipitation (TCP) in Eastern Mexico

Laiyin Zhu[1,1] and Pascual Aguilera[1,1]

[1]Western Michigan University

November 30, 2022

Abstract

Tropical Cyclone Precipitation (TCP) is one of the major triggers of flash flooding and landslide in eastern Mexico. The interactions between the topography of the Sierra Madre Occidental and the TCP of storms from the Gulf of Mexico are still poorly understood. We apply multiple statistical techniques to a 99 year daily TCP record and an elevation data with high spatial resolution. Correlation analysis for the whole dataset is dominated by the strong inland-to-ocean gradient of both TCP and topography. Clusters defined by grids' distances to the coast show significant positive correlations between TCP variables and topographic complexity variables (Range, Standard Deviation, and Slope). The quantile analysis demonstrates that the most extreme TCPs are more likely to locate in grids with higher amounts of topographic complexity (Range and Standard Deviation) than the median and the trivial TCPs. The Random Forest (RF) model is an excellent tool to disentangle complex relationships between TCP and topography. The models show that the grid's location and aspect of the slope aspect are the two most important variables that affect the TCP statistics. TCP in eastern Mexico is sensitive within two zones: (1) Low lying coastal regions with lower elevation and less topographic complexity. (2) The mountainous region with higher elevation and topographic complexity, especially with the slope facing the windward direction to the Gulf. All results support that the topography in eastern Mexico has an enhancing effect on the TCP.

Evaluating Variations in Tropical Cyclone Precipitation (TCP) in Eastern Mexico using Machine Learning Techniques

L. Zhu[1], P. G. Aguilera[2]

[1] Department of Geography, Environment, and Tourism, Western Michigan University, Kalamazoo, MI, USA.

[2] Department of Physics, Western Michigan University, Kalamazoo, MI, USA.

Corresponding author: Laiyin Zhu (laiyin.zhu@wmich.edu)

Supplement 1. List of variables used in the Random Forest Modeling

#	Variable Name	Short Description
1	AMTCP	Annual Mean TCP at each grid
2	MaxETCP	Historical Maximum Event TCP at each grid
3	ETCP	Event TCPs at each grid
4	ETCP90	All TCP events with precipitation greater than the 90 percentile of the ETCP.
5	Mean	Mean elevation within each 0.25° grid box
6	Max	Maximum elevation within each 0.25° grid box
7	Min	Minimum elevation within each 0.25° grid box
8	Range	Difference between Max and Min

#	Variable Name	Short Description
9	StanDev	Standard Deviation for elevations within each 0.25° grid box
10	Slope	The ratio between the rise and the run (tan ϑ) for each 0.25° grid box
11	Aspect	The Slope's major orientation angle from the normal (north as 0°) for each 0.25° grid
12	Distant to Track	Nearest sphere distance from each precipitation grid to the storm track
13	Track Cluster	Track Cluster defined by the regression mixture model by Gaffney et al. (2007)
14	Lon	Longitude of each 0.25° grid
15	Lat	Latitude of each 0.25° grid
16	Distance to Coast	Nearest sphere distance from each precipitation grid to the Gulf of Mexico Coast
17	Month	Month when each TCP event happened
18	Forward U Speed	A vector for the mean of east-west (east as positive sign) component of the storm mo
19	Forward V Speed	A vector for mean of the north-south (north as positive sign) component of the storm
20	Forward Speed	The magnitude of vector U plus vector V
21	Forward Speed Variance	Variance of the Forward Speed for each TCP event
22	Forward Speed Angle	The direction of the Forward Speed measured clockwise starting from the north for e
23	Forward Speed Angle Variance	Variance of the Forward Speed Angle for each TCP event
24	Stalled	A dummy variable that indicates whether the storm is stalled or not (stalled storms r
25	Event Duration	How long a TCP event last
26	ATCP Cluster	The clusters of TCP grids defined by the anomaly of their annual mean TCP

Supplement 2a. Correlation between the Event TCP and Environmental Variables for Cluster 2 TCP Grids

Track Cluster	Distance to Coast	Lon	Lat	Mean	Max	Min	Range	Std	Slope	Aspect
1	0.00*	0.05*	-0.15*	-0.17*	-0.07*	-0.23*	0.20*	0.18*	0.19*	0.05*
2	-0.07*	0.11*	-0.03*	-0.08*	-0.04*	-0.10*	0.04*	0.03*	0.03*	-0.01
3	0.01	-0.14	0.00*	0.06*	0.08*	0.03*	0.09*	0.08*	0.07*	0.01*

* indicates correlation with p<0.01

Supplement 2b. Correlation between the Event TCP and Track Variables for Cluster 2 TCP Grids

Track Cluster	Distance to Track	Forward U Speed	Forward V Speed	Forward Speed	Forward Speed Variance	Fo
1	-0.47*	0.07*	-0.13*	-0.14*	0.02*	0.
2	-0.43*	-0.03*	-0.16*	-0.09*	-0.04	0.
3	-0.45*	-0.09*	-0.24*	-0.07*	-0.02*	0.

* indicates correlation with p<0.01

Supplement 3a. Correlation between the Event TCP and Environmental Variables for Cluster 1 TCP Grids

Track Cluster	Distance to Coast	Lon	Lat	Mean	Max	Min	Range	Std	Slope	Aspect
1	-0.16*	0.21*	-0.14*	-0.24*	-0.16*	-0.26*	0.12*	0.12*	0.13*	0.00
2	-0.07*	0.15*	-0.11*	-0.10*	-0.03*	-0.14*	0.09*	0.06*	0.06*	0.01
3	-0.01	-0.08*	-0.06*	0.06*	0.08*	0.03*	0.09*	0.08*	0.07*	0.01

* indicates correlation with p<0.01

Supplement 3b. Correlation between the Event TCP and Track Variables for Cluster 1 TCP Grids

Track Cluster	Distance to Track	Forward U Speed	Forward V Speed	Forward Speed	Forward Speed Variance	Fo
1	-0.37*	-0.08*	-0.01	0.04*	0.01	0.
2	-0.45*	-0.17*	-0.05*	0.04*	0.01	0.
3	-0.43*	-0.18*	-0.21*	-0.05*	-0.06*	0.

* indicates correlation with p<0.01

Supplement 4. The Variation of Model Performance for the AMTCP from all possible combinations of variable subsets, calculated by the Recursive Feature Selection (RFE) algorithm using three repeated 10 folds cross validations.

Number of Variables	RMSE	R^2	MAE	Selected
1	5.74	0.94	3.08	
2	4.25	0.97	1.89	
3	3.57	0.98	1.39	*
4	3.80	0.97	1.54	
5	4.36	0.96	1.80	
6	4.08	0.97	1.55	
7	4.29	0.96	1.66	
8	4.49	0.96	1.79	
9	4.34	0.96	1.66	
10	4.48	0.96	1.73	

Supplement 5. The Variation of Model Performance for the MAXETP from all possible combinations of variable subsets, calculated by the Recursive Feature Selection (RFE) algorithm using three repeated 10 folds cross validations.

Number of Variables	RMSE	R^2	MAE	Selected
1	90.08	0.33	62.57	
2	40.82	0.86	18.93	
3	39.88	0.87	18.60	*
4	40.77	0.86	19.64	
5	42.36	0.85	20.75	
6	41.18	0.86	19.66	
7	41.89	0.86	20.24	
8	42.37	0.85	20.63	
9	41.83	0.86	20.11	
10	42.40	0.85	20.43	

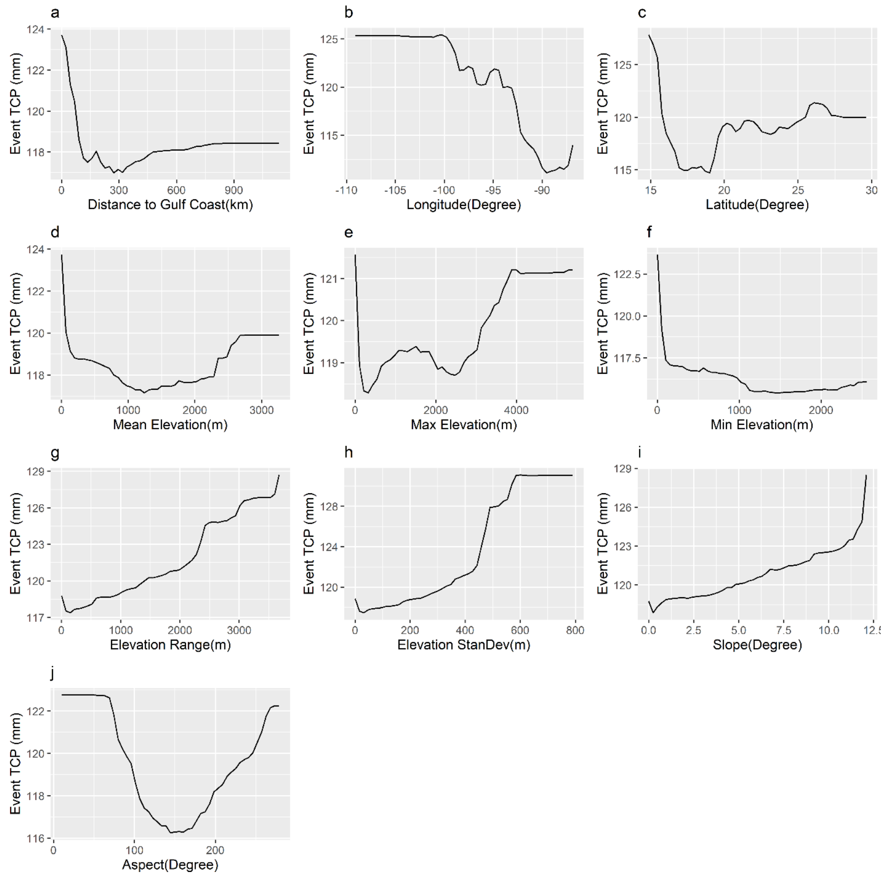
Supplement 6. The Variation of Model Performance for the ETCP from all possible combinations of variable subsets, calculated by the Recursive Feature Selection (RFE) algorithm using three repeated 10 folds cross validations.

Variables	RMSE	R^2	MAE	Selected
1	35.73	0.12	21.57	

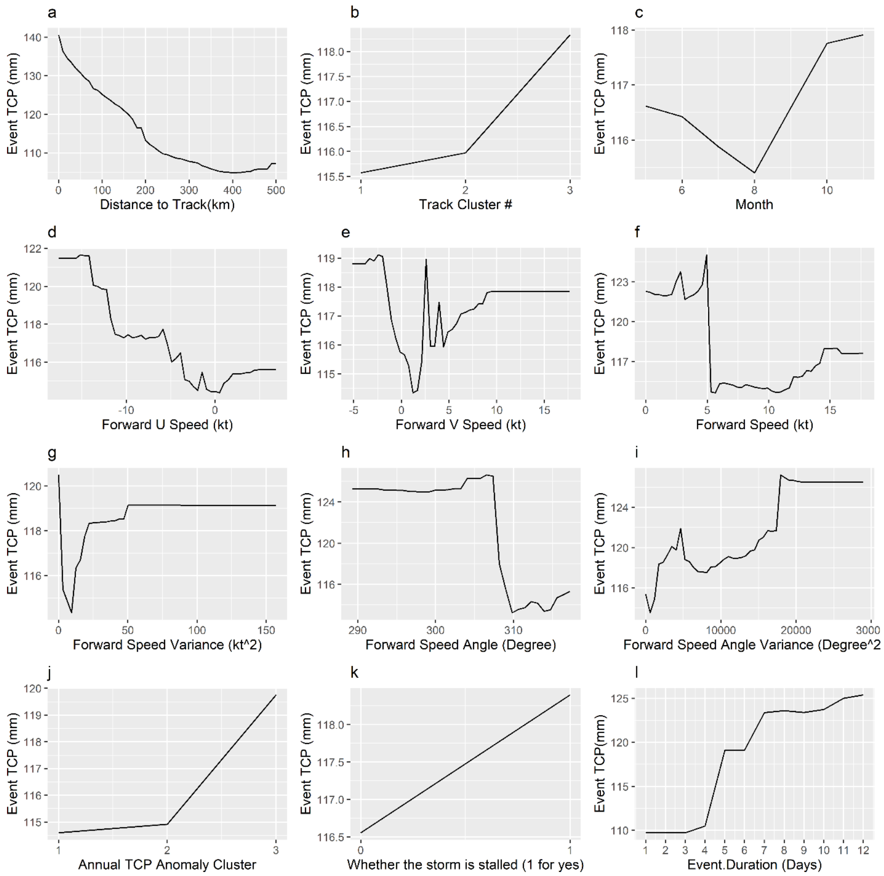
Variables	RMSE	R^2	MAE	Selected
2	32.79	0.21	19.55	
3	23.63	0.61	13.52	
4	21.64	0.69	12.27	
5	17.07	0.82	8.95	
6	14.17	0.86	6.66	
7	14.88	0.85	7.06	
8	15.49	0.83	7.45	
9	14.34	0.85	6.73	
10	14.69	0.85	6.96	
11	14.93	0.84	7.12	
12	14.13	0.86	6.63	
13	14.25	0.86	6.72	
14	14.30	0.86	6.76	
15	13.88	0.86	6.52	
16	13.92	0.86	6.57	
17	13.71	0.87	6.45	
18	13.51	0.87	6.36	*
19	13.51	0.87	6.39	
20	13.52	0.87	6.41	
21	13.52	0.87	6.41	
22	13.56	0.87	6.43	

Supplement 7. The Variation of Model Performance for the ETCP90 from all possible combinations of variable subsets, calculated by the Recursive Feature Selection (RFE) algorithm using three repeated 10 folds cross validations.

Variables	RMSE	R^2	MAE	Selected
1	54.90	0.01	37.56	
2	53.05	0.03	36.22	
3	51.58	0.08	34.92	
4	51.30	0.10	34.62	
5	46.36	0.27	30.94	
6	37.87	0.51	24.69	
7	36.15	0.56	23.47	
8	34.33	0.62	22.15	
9	32.78	0.64	20.97	
10	32.75	0.64	20.92	
11	32.85	0.64	20.99	
12	32.56	0.65	20.80	*
13	32.81	0.64	21.02	
14	33.04	0.64	21.21	
15	32.90	0.64	21.13	
16	33.08	0.64	21.28	
17	33.22	0.63	21.38	
18	33.05	0.64	21.26	
19	33.08	0.63	21.28	
20	33.08	0.63	21.28	
21	32.92	0.64	21.16	
22	32.97	0.64	21.19	



Supplement 8. Partial Dependence Plot for static variables in the Whole Model for the ETCP90



Supplement 9. Partial Dependence Plot for dynamic variables in the Whole Model for the ETCP90

Supplement 10. Comparison of the median values for the extreme (> 90th percentile, P90) TCP and the median range (between 45th percentile and 55th percentile) TCP samples

Storm	Hurricane Alex	Hurricane Alex	Hurricane Igrid	Hurricane Igrid	Hurricane Beulah	Hurricane Beulah
Quantile	P_{90} TCP	P_{45} to P_{55} TCP	P_{90} TCP	P_{45} to P_{55} TCP	P_{90} TCP	P_{45} to P_{55} TCP
Cluster 1	71.03	25.83	52.40	22.06	91.53	37.10
Cluster 2	167.55	78.70	208.99	65.67	205.19	60.04
Cluster 3	171.58	70.55	129.98	55.02	292.67	113.44

*Clusters are defined by K-Means of the Annual Mean TCP Anomaly.

Supplement 11. Comparison of median of elevation standard deviation for the extreme (> 90th percentile, P_{90}) TCP and the median range (between 45th percentile and 55th percentile, P_{50}) TCP samples

Storm	Hurricane Alex	Hurricane Alex	Hurricane Igrid	Hurricane Igrid
Quantile	P_{90} TCP Elevation Std	P_{45} to P_{55} TCP Elevation Std	P_{90} TCP Elevation Std	P_{45} to P_{55} TCP Elevation
Cluster 1	190.85	144.13	674	630
Cluster 2	22.37	18.98	1294*	512
Cluster 3	172.08	204.09	275*	115

*Clusters are defined by K-Means of the Annual Mean TCP Anomaly.

Supplement 12. Comparison of the median elevation range for the extreme (> 90th percentile, P_{90}) TCP and the median range (between 45th percentile and 55th percentile, P_{50}) TCP samples

Storm	Hurricane Alex	Hurricane Alex	Hurricane Igrid	Hurricane Igrid
Quantile	P_{90} TCP Elevation Range	P_{45} to P_{55} TCP Elevation Range	P_{90} TCP Elevation Range	P_{45} to P_{55} TCP E
Cluster 1	972	926.5	674	630
Cluster 2	127	130	1294*	512
Cluster 3	1102.5	1146	1807*	903.5

*Clusters are defined by K-Means of the Annual Mean TCP Anomaly.

Evaluating Variations in Tropical Cyclone Precipitation (TCP) in Eastern Mexico using Machine Learning Techniques

L. Zhu[1], P. G. Aguilera[2]

[1] Department of Geography, Environment, and Tourism, Western Michigan University, Kalamazoo, MI, USA.

[2] Department of Physics, Western Michigan University, Kalamazoo, MI, USA.

Corresponding author: Laiyin Zhu (laiyin.zhu@wmich.edu)

Key Points:

- Tropical Cyclone Precipitation (TCP) variations are evaluated using statistical and machine learning methods based on a 99-year climatology.
- The RF model has an excellent fitting and predicting skill in TCP, and it captures complex and nonlinear relationships controlling the TCP.
- The annual mean TCP is determined by locations, while the event TCP is determined by interactions of multiple dynamic and static variables.

Abstract

Tropical Cyclone Precipitation (TCP) is one of the major triggers of flash flooding and landslide in eastern Mexico. We apply different statistical and machine learning techniques to study a 99 year TCP climatology in high resolution. Strong correlations exist between location variables and annual mean TCP, as well as between dynamic variables and event TCP. Topographic variables observe mixed signals with the elevation variances positively correlated with TCP. The Random Forest (RF) model is a powerful tool with excellent fitting and predicting skills for TCP variations. It has a very small out of sample cross-validation error and well captures the spatial variations of historical TCP events. Only three location variables are needed to construct the best model for the annual mean TCP while the best model needs 18 variables to explain the complex variations in the event TCP. The distance to the track is the most important variable for the event TCP model and many other factors contribute to the TCP collectively and nonlinearly, which can’t be captured fully by the previous correlation analysis. They include translation characteristics of the storms, locations of the precipitation grid, and topography. Event TCP is generally larger in storms with slower translation speed and more variance in their tracks. While the lower coastal area generally has a higher probability of TCP, the higher inland has elevation variances that enhance less frequent but extreme TCP events. The RF algorithm is an efficient machine learning approach showing potentials for future Quantitative Precipitation Forecasting (QPF).

1 Introduction

Tropical Cyclone Precipitation (TCP) is one of the major triggers of flooding and landside. The TCP processes are complex and influenced by many factors, which include the moisture and energy that the storm brought from the ocean, the shape and size (Matyas, 2007; Zhou et al., 2018), the translation speed, the intensity of the storm, the surface conditions of the land (moisture and energy), land use and cover, interactions with other weather systems, and the topographic features (Arndt et al., 2009; Kimball, 2008; Tuleya, 1994; Zhang et al., 2018). Different studies (Emanuel, 2017; Knutson et al., 2019; Risser & Wehner, 2017; Trenberth et al., 2018) have argued that anthropogenic global warming may increase the chance of extreme TCP events like Hurricane Harvey in 2017 and the majority of the modeling community holds high or medium-to-high confidence that the rain rate for TCs is going to increase by 14% with 2°C of warming (Knutson et al., 2020). This is consistent with the Clausius-Clapeyron equation. TCP over the land has high spatial variability (Skok et al., 2013; Zhu & Quiring, 2013). TC track is an important factor controlling the storm precipitation. Slower moving storms are contributing to more local rainfalls with longer duration of rain events and possibly higher rain rates (Chan, 2019; Kossin, 2018). The boundary layer condition is significantly changed when TCs make landfall. Increases in land surface roughness can enhance topographic advection (Arndt et al., 2009; Kimball, 2008; Tuleya, 1994; Zhang et al., 2018) and introduce more TCP by influencing the low-level convergence (Kepert, 2001; Langousis & Veneziano, 2009; Shapiro, 1983). Many modeling and observation studies proved that topography has an enhancing effect on TCP (Huang et al., 2020; Li et al., 2007; Ramsay & Leslie, 2008; Wu et al., 2002) based on different dynamic processes. Houze (2012) provided a physical mechanism for the lifting effect of tropical cyclones by the topography. While TCs are over the ocean they tend to be moist neutral and the uniform warm ocean boundary makes the flow slightly unstable. The lifting over the mountainside releases this instability and triggers the convective cells on the windward side and then interacts with the gravity wave on the lee side of the mountain. Sometimes the TCP process is further complicated by the interactions of the storm track, land/ocean distributions, and topography over the land. Topography has been reported to deflect TC tracks and change their precipitation intensity over the land (Huang et al., 2012; Lin et al., 2005; Lin et al., 2002).

Mexico is a country with a complex topography and long coastal lines prone to TCs on both sides. Existing works on precipitation in Mexico are focused on general precipitation (Mascaro et al., 2014; Pineda-Martinez & Carbajal, 2009), North American Monsoon (Vivoni et al., 2007) and TCP mechanisms on the Pacific Coast (Farfán & Cortez, 2005; Farfán & Zehnder, 2001; Zehnder, 1993). TCP can contribute 0 to 40% of the annual precipitation across Mexico, which is estimated from the satellite precipitation product TMPA 3B42 from 1998 to 2013 (Agustín Breña-Naranjo et al., 2015). Franco-Díaz et al. (2019) used the same product and estimated that TCs contribute 10 to 30% of July to October precipitation and they are associated with 40 to 60% of coastal daily extreme rainfall (> 95th percentile) in Mexico. Extreme TCP events in Mexico are triggers of severe flooding with massive disruption to society and intense economic losses (Agustín Breña-Naranjo et al., 2015). Two TCs (Tropical Storm Manuel and Hurricane Ingrid) made landfall in Mexico between September 13 and 20 in 2013. Flooding from extreme precipitation has damaged 45000 homes with $900 million of insured losses and $5.7 billion in total economic losses. Therefore, it is necessary to systematically evaluate the variations of the TCP on the east side of Mexico and the factors that influence it. Our analysis is based on a 99-year daily gridded TCP record derived from a large number of rain gauges. It is possibly the longest climatological record that can be discovered for the region with acceptable details. We will evaluate the relationships by using multiple statistical and data mining techniques including cluster analysis, correlations, and the Random Forest (RF) models. We will develop the optimal Random Forest models for variations in both annual mean and event TCP and evaluate their fitting and predicting skills from out-of-sample cross-validations.

The article is organized as follows. Section 2 will introduce the data and methods of the analyses with more details. In Section 3, we will present the results from different statistical and data mining methods and a case study focused on the three most extreme historical events. We will summarize and discuss our findings in Section 4.

2 Data and Methods

2.1. Precipitation

The TCP is extracted from daily rain gauges and locations of the TC for both the U.S. and Mexico from 1920 to 2018. The Daily Global Historical Climatology Network (GHCN-D) covers both the U.S and Mexico with 35161 gauges. The GHCN-D has decent spatial density for spatial interpolation into 0.25° grids inside the U.S. but is not dense enough for Mexico. Therefore, we collect a second source of daily precipitation from 2526 gauges provided by the National Weather Service of Mexico. We define daily TCP boundaries by connecting moving circles with a radius of 800 km, which are centered by the 6-hour locations provided by the International Best Track Archive for Climate Stewardship (IBTrACS). We use the same approach as Zhu and Quiring (2017), which gives the optimal estimation of 0.25° gridded TCP by correcting possible wind introduced under-catches in rain gauges and optimizing the Inverse Distance Weighting (IDW) parameters for the spatial interpolation. The algorithm was validated with the Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis product 3B42 (TMPA 3B42). The daily TCP grids are then clipped by daily boundaries defined by the connected 500 km radii. The 500 km radii are the final boundaries of the daily TCP and the previous 800 km circles are used to avoid bias in the IDW spatial interpolation, particularly near the 500 km boundary edges. We have identified 4373 TCP days for the whole North American Continent and 1442 TCP days for Mexico between 1920 and 2018. Figure 1a shows that we have enough rain gauge density in the study are for the IDW algorithm: the numbers of gauges are far more than the final interpolated grids in eight decades after 1940. The decade with the lowest number of gauges is 1920 to 1929, which still has an average gauge/grid ratio of greater than 1/2.

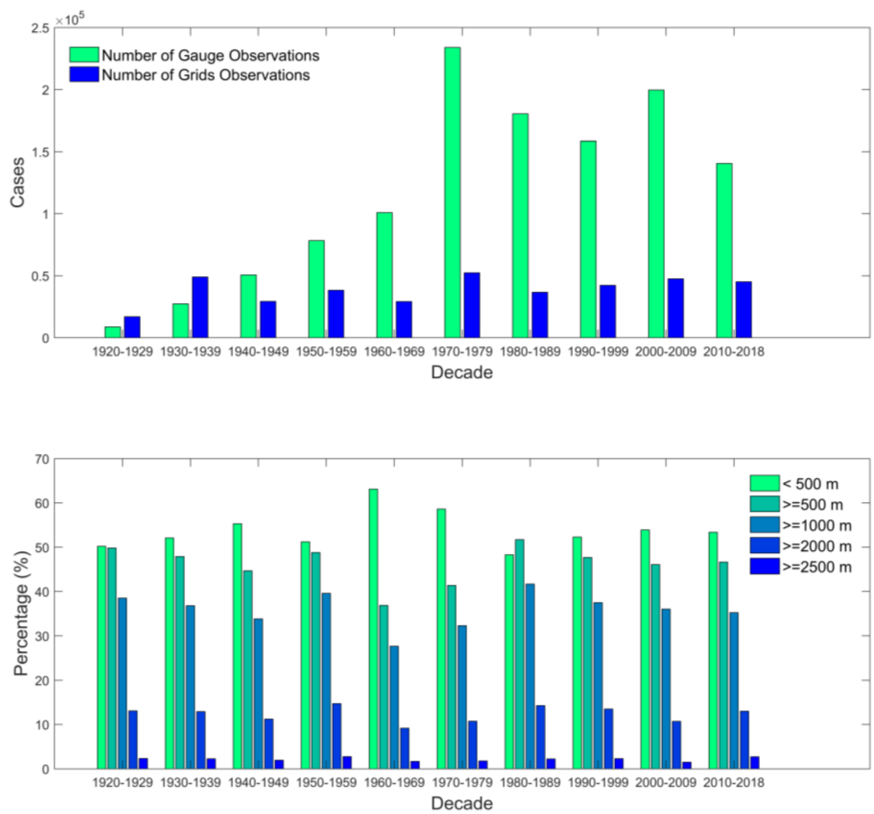


Figure 1. Statistics for (a) the total number of gauges and interpolated grids (0.25°) for daily TCP (b) percentage of grids in different elevation ranges.

The daily TCPs are also aggregated into storm total TCP, which yields 399 TCP events. Annual Mean TCP, Maximum Event TCP, and the 90th Percentile (P_{90}) TCP are also calculated for comparison and modeling purposes. Because there is a generally decreasing gradient of TCP probability from the coast locations to the inland locations, we define three clustered regions of our grids based on their annual TCP anomaly (Figure 2) using the K-Means clustering method. The reason is that variables that influence the TCP are also determined by their locations. One case is that the topography also has the coast-to-inland gradient. The three clusters demonstrate a clear separation pattern from coast to inland and they will be used in the subsequent correlation analysis and RF modeling.

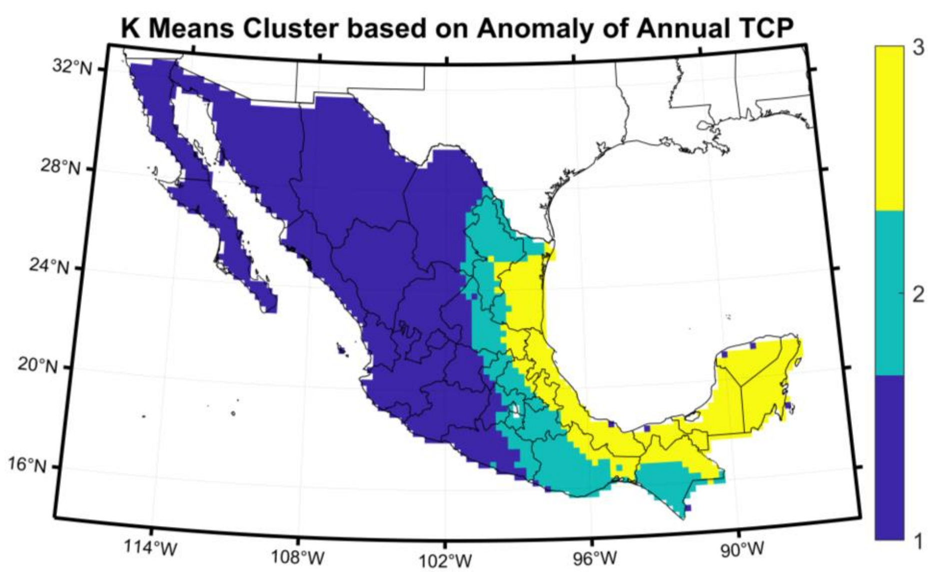


Figure 2. K-Means Clusters of grids calculated based on their annual mean TCP anomaly.

2.2. Topography and Location Variables

We obtain the raw elevation data from the Global 30 Arc-Second Elevation (GTOPO30) offered by the Earth Resources Observation and Science (EROS) Center of the United States Geological Survey. The GTOPO30 has a 1 km resolution and was derived from a variety of sources in 1996. We calculate seven elevation variables from ~ 750 GTOPO30 points within each 0.25° grid box. We estimate the mean, maximum, minimum, and standard deviation of the elevations for each box. The range is defined as the difference between the highest and the lowest elevation inside each box. The slope and its' aspect are calculated by the algorithm (Burrough et al., 2015) provided by the ESRI ArcGIS zonal statistics package. The slope is the mean steepness for each 0.25° box and the aspect is the slope's direction measured clockwise from 0° (due north). We will analyze how those topographic variables are related to the TCP. Figure 1b also shows that we have decent amounts of grids within each elevation range for all ten decades, which adds confidence to our subsequent data analysis for the elevation and TCP. We also calculate the centroid longitude and latitude for each 0.25° grid and the sphere distance from each centroid to the nearest coastline of the Gulf of Mexico (distance to the coast) because they may all influence the spatial variations of TCP.

2.3. TC Tracks and Characteristics

TC track characteristics are important factors that determine the amount of individual storm precipitation. Here we take all TC track sections (locations recorded at 6-hour intervals) that impacted Mexico with precipitation (the parts of tracks overland or near the land) and define them into 3 different clusters using the storm track clustering technique developed by Gaffney et al. (2007). This clustering technique uses the functions of the cyclone positions conditioned on an independent variable time as the conditional density components for the regression mixture model framework (Camargo et al., 2007). Details for the algorithm can be found from the Matlab toolbox that is freely available at http://www.datalab.uci.edu/resources/CCT. Figure 2 shows that those clusters have different spatial patterns. The cluster 1 tracks are more located in the south part of Mexico with a curve feature for their cluster mean track. The cluster 2 tracks are more likely to penetrate through Mexico in the middle. The cluster 3 tracks are more located in the northern part of Mexico bordered to Texas, U.S with a curve feature as well. We will use these track clusters in our following analysis.

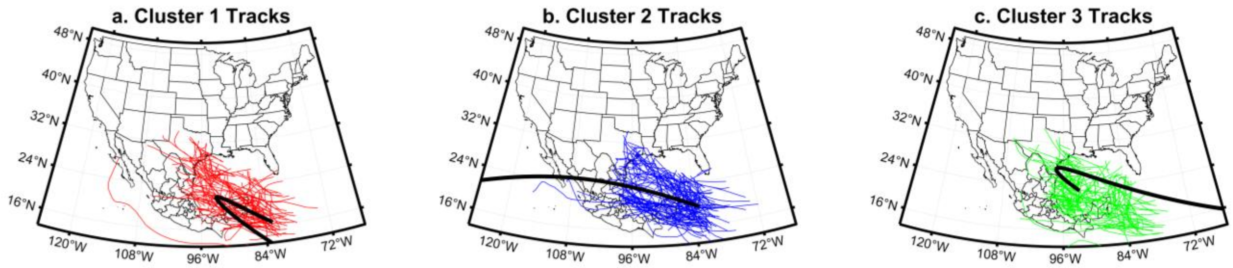


Figure 3. Clusters of TC tracks for storms generating precipitation in Mexico, colored lines are actual TC tracks, and the black line is the cluster mean track estimated by the model.

In addition to the spatial clustering of tracks, other TC properties may also determine the amount of TCP in each event. We calculate several different properties for all 399 events. The distance to track is defined as the closest sphere distance (km) between each precipitation grid and the storm track. The forward U speed (kt) is defined as a vector of the mean of the east-west (east as positive sign) component of the storm movement, while the forward V speed (kt) is the vector for the mean of the north-south (north as positive sign) component of the storm movement. The forward speed (kt) is the magnitude of vector U plus vector V, and the forward speed angle is the direction of the forward speed measured in degrees clockwise from the north. We also calculate the variances for both the forward speed and its angle along each of the storm track to capture changes in its movement. We define a dummy variable that indicates whether the storm is stalled or not (stalled storms are defined as '1' if they ever moved toward the south while other storms are defined as 0). Finally, we also calculate the event durations by summing all TCP days for each event.

2.4. Data Analysis and Model Development

We apply the pairwise correlations (Spearman's ρ) with p-values (<0.01) (Best & Roberts, 1975) to explore the relationships between the TCP and factors that may influence it. We also apply percentile analysis to compare samples in the TCP data using the Mann-Whitney U-test (Mann & Whitney, 1947) to compare the sample mean of elevation characteristics for different TCP groups. Traditional statistical techniques like correlation or linear regressions are straightforward for the interpretation of the signals. However, they lack the ability in capturing combined effects from multiple independent variables and nonlinear relationships, as well as suffer issues like collinearity. And they are not able to deal well with variables with specialized distributions (e.g., slope aspect with a cyclic change from 0 to 360°).

The RF model is a powerful machine learning algorithm (Breiman, 2001; Breiman et al., 1984) with a much less stringent requirement for distribution or type of independent variables. The algorithm fits a large number (K=500 in our study) of regression trees by using bootstrapped training samples. The data are recursively partitioned into two groups based on a subset of explanatory variables in each tree until the terminal nodes reach minimum size. The model prediction is based on the ensemble of K regression trees. The randomness in both the bootstrap sampling and the selection of subset predictors at each node of the trees results in the reduction of the correlation between trees (Nateghi et al., 2014). The RF algorithm is easy to implement. It can capture the complex nonlinear feature of the data and offer excellent prediction accuracy. The TCP is a complex process determined by multiple factors together and many of those variables are not normally distributed. We believe that the RF algorithm is an excellent candidate to explore those relationships and can potentially yield powerful prediction models.

We will develop two sets of RF models for TCP in Mexico, using the TCP metrics and explanatory variables we developed in sections 2.1 to 2.3. A detailed list of all dependent and independent variables can be found in Supplement 1. The first set of models are focused on the aggregated TCP statistics for the entire 99 years. We will model the Annual Mean TCP (AMTCP) and Historical Maximum Event TCP (MAXETCP) at each grid. The independent variables are all static (Variable # 5-11, 14-16 in Supplement 1). The second set of models are focused on event TCP (ETCP) and > P_{90} event TCP (ETCP90), which are developed from both static and dynamic independent variables totaled by 22.

Samples for both AMTCP and MAXETCP contain 2775 records. The ETCP sample has 165667 records and the ETCP90 sample has 16567 records. Because of the large data volume, both ETCP and ETCP90 models are trained and validated by using the High-Performance Computing (HPC) facility (Pitzer Clusters from the Ohio Supercomputer Center). We develop two models for each of the four dependent variables: (1) a whole model that includes all explanatory variables and all data. (2) a "best" model that uses the Recursive Feature Elimination algorithm to select an optimal subset of explanatory variables that gives the best cross-validation result in out of sample prediction. The whole model (1) is developed to show the partial dependence plots (pdp) for all explanatory variables. The pdp explains the marginal effect of each explanatory variable on the response variable while effects from other explanatory variables are averaged out (Hastie et al., 2009). It is an effective tool to explain the contribution from each explanatory variable by capture its variability and particularly the non-linear relationships with the dependent variable. The R package for the pdp is freely available from the internet (https://cran.r-project.org/web/packages/pdp/). The best model (2) is developed for the best cross-validation performance, we separate the whole sample into 80% training data and 20% testing data. Then we use the "caret" R package (available at https://cran.r-project.org/web/packages/caret/) to train our RF models. The model is trained by using the repeated cross-validation approach, which randomly selects 10 folds of the training data to construct the model and use the remaining of training data to validate the model. And this process is repeated three times and all error statistics are summarized. We use the Recursive Feature Selection (RFE) function to choose the optimal subset of variables to be included in the final model by testing all possible combinations of variables. The criteria for final model selection is based on the ensemble mean Root Mean Squared Error (RMSE). We also use the best model to make predictions for the 20% testing data that has not participated in the model fitting. We will report performance statistics for the 20% testing sample, the 80% training sample (fro repeated cross-validation), and the whole sample. Those performance statistics include the RMSE, the Mean Absolute Error (MAE), and R^2. The RF model can give the value and rank of the Variable Importance (VI) in the model and reveal relationships and sensitivities between independent variables and response variables (Greenwell, 2017). The VI is computed as the usefulness of each independent variable in splitting the data at each node of the regression tree and a "pure" node is preferred. The VI is measured by the increase of Gini impurity, calculated based on the reduction in the sum of squared errors whenever a variable is chosen to split (Strobl et al., 2007). We then normalize the VI based on a 0-100 scale for easier comparison across models (McRoberts et al., 2018).

3. Results

3.1. Spatial Patterns and Summary Statistics

Figure 4 shows the maps for the mean elevation, elevation range, AMTCP and MAXETCP for Mexico. Mexico has mountainous areas higher than 3000 meters in the central and areas below 500 meters on the coast (Figure 4a). Transition zones with large elevation changes (range) are located between the coast and inland area (Figure 4b). The AMTCP (Figure 4c) shows a strong decreasing gradient from the coast to inland. This gradient still exists for the

MAXETCP (Figure 4d) but not as strong as AMTCP. The MAXETCP also has scattered local maximums over inland locations, which may indicate the topographic enhancement of TCP.

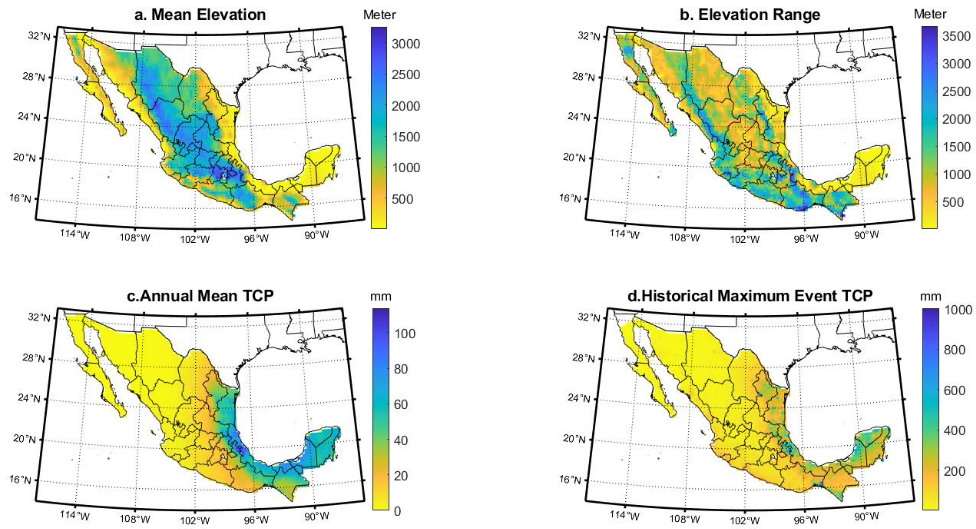


Figure 4. Spatial patterns in Elevation (Mean Elevation and Elevation Range) and TCP characteristics (AMTCP and MAXETCP) in Mexico

Correlations between environment variables and AMTCP and MAXETCP are shown in Table 1. Both AMTCP and MAXETCP are most sensitive to location variables and they show the strongest correlations. Higher TCP generally corresponds to locations nearer the coast, as well as at more eastern and southern positions. The elevation variables are showing mixed results. For cluster 1 locations in more mountainous areas, the elevation variables are demonstrating more positive correlations with the TCP, which again indicates the enhancing effect of TCP from the topography. However, cluster 2 and particularly cluster 3 locations are showing negative correlations for many elevation variables. The distance to the coast also determines spatial changes of elevation. Coastal areas are mostly associated with lower elevations but have a higher general probability of TCP. The correlations in aspect are hard to interpret because of their cyclic distribution.

Table 1. Correlation between TCP variables and Environmental Variables.

Var	Cluster	Distance to Coast	Lon	Lat	Mean	Max	Min	Range	Std	Slope	Aspect
AMT	1	-0.74*	0.76*	-0.61*	0.24*	0.23*	0.24*	0.05	0.11*	0.09*	-0.13*
CP	2	-0.51*	0.09	-0.16*	-0.14*	-0.15*	-0.17*	-0.11*	-0.10	-0.12*	0.13*
	3	-0.74*	0.45*	-0.05	-0.24*	-0.17*	-0.30*	0.04	0.02	0.05	-0.12
MA	1	-0.54*	0.59*	-0.62*	-0.06*	0.04	-0.11*	0.21*	0.22*	0.28*	-0.02
XET	2	-0.27*	0.16*	0.06	-0.06	-0.10	-0.06	-0.09	-0.06	-0.06	0.08
CP	3	-0.25*	0.26*	0.19*	-0.55*	-0.41*	-0.55*	-0.04	-0.07	-0.07	-0.21*

* indicates correlation with p<0.01, Clusters are defined by K-Means of the AMTCP anomaly in Figure 2

We also conduct correlation analyses between event TCP (ETCP) and selected explanatory variables. The ETCP contains 165667 observations and has far more variance than the aggregated records (AMTCP and MAXETCP) so we expect more complex relationships. Here we show an example of correlations for cluster 1 grids in table 2a and 2b, results for the other two clusters are demonstrated in supplement 2 and 3. Because of the much larger sample size, most of the correlations are significant with $p<0.01$. The distance to coast, longitude, and latitude have a similar relationship with the ETCP as they have with the AMTCP (Table 1), but with lower correlation values. The mean, max, and min elevation are showing negative correlations with the ETCP for storms with cluster 1 and 2 tracks, but they have positive correlations for storms with cluster 3 tracks. Storms with cluster 3 tracks tend to make landfall in northern Mexico, and the elevation is relatively higher there and possibly enhance the TCP. The range, standard deviation and slope are all showing positive correlations with the TCP for all track clusters, which demonstrates that the elevation variances have consistent positive contributions to more TCP. If we look at the track variables in Table 2b, the distance to track has the strongest negative correlation with ETCP among all variables. It also generally shows that the slower-moving storms are generating more ETCP. This relationship is particularly strong for the north-south direction (forward V speed) of storm movement. Those relationships are similar for Cluster 2 and 3 grids (Supplement 2 and 3) with some variations.

Table 2a. Correlations between the Event TCP and Static Variables for Cluster 1 TCP Grids

Track Cluster	Distance to Coast	Lon	Lat	Mean	Max	Min	Range	Std	Slope	Aspect
1	-0.19*	0.22*	-0.23*	-0.14*	-0.09*	-0.18*	0.11*	0.11*	0.12*	0.03*
2	-0.07*	0.12*	-0.08*	-0.07*	-0.02*	-0.11*	0.07*	0.06*	0.06*	0.01
3	0.00	-0.12*	-0.02*	0.05*	0.08*	0.02*	0.09*	0.08*	0.07*	0.02*

* indicates a correlation with $p<0.01$

Table 2b. Correlations between the Event TCP and Track Variables for Cluster 1 TCP Grids

Track Cluster	Distance to Track	Forward U Speed	Forward V Speed	Forward Speed	Forward Speed Variance	Forward Speed Angle	Forward Speed Angle Variance
1	-0.36*	-0.07*	-0.05*	0.05*	-0.02*	0.18*	-0.01*
2	-0.41*	-0.11*	-0.22*	0.04*	0.00	0.18*	-0.10*
3	-0.44*	-0.06*	-0.29*	-0.14*	-0.08*	0.14*	-0.14*

* indicates a correlation with $p<0.01$

3.2 Random Forest Model

3.2.1. The AMTCP and MAXETCP

RF models are developed for both AMTCP and MAXETCP using locations and topographic information as independent variables. The RF models show very high fitting and predicting skills for the AMTCP and MAXETCP. The AMTCP models generally have less error and higher R^2 values than the MAXETCP models. The whole models are fitting the entire data better but have worse performance in predicting the subsets of the data (testing and training samples). The best models are trained only from the training sample and have better out of sample performance (testing sample). Interestingly, the AMTCP and MAXETCP best models have only three identical participating variables: distance to coast, longitude, and latitudes. They are all location variables and can explain most of the variance in AMTCP and MAXETCP in Mexico and offer better error statistics than the whole models fitted by 10 Variables.

Table 3. Model Performance Summary for the Whole Model and the Best Model of the AMTCP and the MaxETCP

	AMTCP						MaxETCP					
Model	Whole Model			Best Model			Whole Model			Best Model		
Sample	Test	Train	Whole	Test	Train*	Whole	Test	Train	Whole	Test	Train*	Whole
RMSE	2.13	4.59	1.92	2.09	3.57	2.03	34.28	44.99	22.34	33.84	39.88	26.69
MAE	1.08	1.69	0.71	1.07	1.39	0.86	34.27	20.73	10.11	17.89	18.60	12.39
R^2	0.99	0.96	0.99	0.99	0.98	0.99	0.90	0.83	0.96	0.90	0.87	0.94

***** indicates that statistics are calculated from the RFE multiple cross-validation routine.

Table 4. Variable Importance (VI) Summary for the Whole Model and the Best Model of the AMTCP and the MaxETCP

	AMTCP				MaxETCP			
	Whole Model		Best Model		Whole Model		Best Model	
Rank	Name	VI	Name	VI	Name	VI	Var Name	VI
1	Distance to Coast	100	Distance to Coast	38.28	Distance to Coast	100	Lat	37.11
2	Lon	44.43	Lon	34.75	Lon	66.51	Lon	32.50
3	Lat	8.18	Lat	29.33	Lat	21.75	Distance to Coast	30.13
4	Max	2.04			Min	9.41		
5	Min	1.85			Mean	5.74		
6	Mean	1.18			StanDev	1.40		
7	StanDev	0.56			Max	1.21		
8	Slope	0.21			Aspect	0.52		
9	Range	0.20			Range	0.14		
10	Aspect	0.00			Slope	0.00		

3.2.2. The ETCP and ETCP90

Both the Event TCP (ETCP) and the Event TCP greater than 90 percentile (ETCP90) include more variabilities than the AMTCP and MAXETCP. All storm events vary in their characteristics, such as track, moisture content, interactions with the land surface, etc. Those factors determine how much precipitation they can generate over land. Our ETCP and ETCP90 models are constructed from 22 potential explanatory variables. Their fitting and predicting skills are slightly worse than the AMTCP and MAXETCP models, but they have much higher model complexity and variability. Table 5 shows that the best models have more consistent performance than the whole models, particularly for the testing and training samples. The best model for the ETCP can explain equal or more than 87% of the variance for different data samples with very low RMSE (8.21 to 13.51 mm) and MAE (3.51 to 6.36 mm). The ETCP90 models are constructed for the most extreme TCP and their performances are worse than the ETCP models. However, the best model for the ETCP90 can still explain 65% to 88% of sample variance with 20.22 to 32.48 mm in RMSE and 11.72 to 20.41mm in MAE.

Table 5. Model Performance Summary for the Whole Model and the Best Model of the ETCP and the ETCP90

	ETCP						ETCP90					
Model	Whole Model			Best Model			Whole Model			Best Model		
Sample	Test	Train	Whole	Test	Train*	Whole	Test	Train	Whole	Test	Train*	Whole
RMSE	13.02	14.16	7.87	13.32	13.51	8.21	33.48	34.35	19.92	32.48	32.56	20.22
MAE	6.10	6.77	3.33	6.23	6.36	3.51	20.71	22.20	11.28	20.41	20.80	11.72
R^2	0.88	0.85	0.96	0.87	0.87	0.95	0.63	0.60	0.88	0.66	0.65	0.88

***** indicates that statistics are calculated from the RFE multiple cross-validation routine.

There are 18 variables in the best model for the ETCP, which shows much higher diversity than the only three location variables chosen by the AMTCP best model. The dynamic variables in the ETCP best model include the distance to track (the most important variable to ETCP), six storm translation parameters (e.g,. forward V speed), track cluster, event duration, and month. Those dynamic variables play the most important role in the model and they are showing higher VI in Table 6. Location variables are the second important variable groups. Latitude, longitude, and distance to coast rank second, fourth and 17th respectively in the VI. We also have five topographic variables participating in the best model: aspect, standard deviation, range, slope, and maximum elevation.

Table 6. The Variable Importance (VI) for the Whole Model and the Best Model of the ETCP

	Whole Model		Best Model	
Rank	Name	V	Name	VI
1	Distance to Track	100.00	Distance to Track	100.00
2	Forward V Speed	57.20	Lat	65.51
3	Lon	41.37	Forward V Speed	54.18
4	Lat	30.51	Lon	42.92
5	Forward Speed Angle Variance	26.73	Forward Speed Angle Variance	38.51
6	Forward Speed Variance	26.45	Forward Speed Variance	36.96
7	Distance to Coast	20.22	Forward U Speed	34.67
8	Forward U Speed	17.51	Forward Speed	27.26
9	Forward Speed	17.39	Forward Speed Angle	26.56
10	Forward Speed Angle	17.37	Track Cluster	25.66
11	Event Duration	16.85	Aspect	24.56
12	Min	10.23	Event Duration	24.21
13	Range	6.50	StanDev	22.74
14	Aspect	6.19	Range	22.60

15	Mean	5.97	Month	22.11
16	Slope	5.45	Slope	21.34
17	Month	5.35	Distance to Coast	19.75
18	StanDev	5.20	Max	17.14
19	Max	5.16		
20	ATCP Cluster	4.49		
21	Track Cluster	2.76		
22	Stalled	0.00		

The VI ranking for the ETCP90 models (table 7) is demonstrating some differences from the ETCP models. The best model has 17 variables and they show less difference between each other in their VIs. The dynamic variables and the location variables are still demonstrating their high importance. Elevation variables have higher VIs than they have in ETCP models, indicating that the elevations play more important roles in determining the most extreme precipitation generated by TCs. The minimum, mean elevation, and the slope aspect rank as 4th, 8th, and 10th important variable in the model, respectively.

Table 7. The Variable Importance (VI) for the Whole Model and the Best Model of the ETCP90

	Whole Model		Best Model	
Rank	Name	VI	Name	VI
1	Distance to Track	100.00	Lon	100.00
2	Lon	63.27	Distance to Track	96.10
3	Lat	62.10	Lat	94.42
4	Distance to Coast	38.33	Min	61.97
5	Forward Speed Variance	34.52	Distance to Coast	56.09
6	Aspect	34.28	Forward Speed Angle	55.76
7	Forward Speed Angle	32.60	Forward Speed Variance	53.71

8	Forward V Speed	32.30	Mean	50.78
9	Forward Speed Angle Variance	31.75	Forward Speed Angle Variance	50.62
10	Forward Speed	30.21	Aspect	50.00
11	StanDev	27.06	Event Duration	49.80
12	Range	24.75	Max	49.79
13	Min	24.33	Forward V Speed	48.86
14	Mean	24.02	StanDev	48.80
15	Forward U Speed	21.48	Range	48.64
16	Slope	20.98	Slope	48.16
17	Max	19.22	Forward U Speed	47.22
18	Event Duration	16.22		
19	Track Cluster	5.64		
20	Month	5.33		
21	Stalled	3.47		
22	ATCP Cluster	0.00		

Lastly, although the ECTP best model provides a nice overall prediction accuracy (Figure 5a), the model's skills deteriorate for the most extreme TCP events (> 69.47 mm, P_{90}) shown in Figure 5b. The R^2 changes from 0.95 to 0.85, and the RMSE increases from 8.21 mm to 22.21 mm. The ETCP90 best model is developed only from a much smaller extreme TCP events sample. It has significant improvement in R^2, RMSE and MAE values if compared with the ETCP best model, Figure 5c also shows many of those improvements happen in the range between 70 mm and 300 mm. All best models have small systematic under-prediction bias across all ranges of TCP, the bias are larger in the most extreme TCP events (> 450 mm).

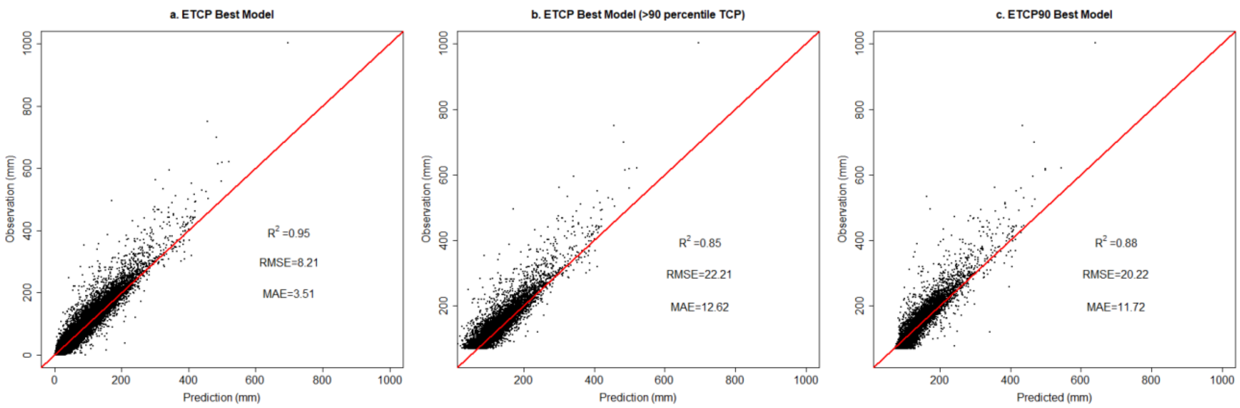


Figure 5. Scatter plots between observation and prediction for the (a) ETCP Best Model for the Whole Sample, (b) ETCP Best Model for the Sample with TCP > 90 percentile, (c) ETCP90 Best Model for the Whole Sample.

3.3 Model Interpretation

Partial dependence plots (pdp) are used to interpret the marginal contribution of each explanatory variable to the response variable of the RF model with the remaining explanatory variables averaged out. We can observe the response variable changes as a continuous function of each explanatory variable independently. This is particularly useful in interpreting the nonlinear relationships inside a complex RF model. We display the pdps of the whole model for both ETCP (Figure 6 and 7) and ETCP90 (Supplement 8 and 9) and they both include all 22 potential explanatory variables. Those 22 variables can be separated into static variables and dynamic variables. The ETCP generally drops when the distance to the coast is less than 400 km but slightly increases when it is between 500 to 1000 km (Figure 6a). The ETCP is generally higher when the longitude is changing from -110° to -95° (Figure 6b), which represents the increase of TCP from the inland to coast (west to east). After a dip, the TCP increases again when longitude is more eastern than -91°, which reflects the TCP received by the Yucatan Peninsular in the most east side of Mexico. The ECTP has the most sensitivity with the latitude (Figure 6c) among all 10 static variables. The TCP generally decreases when the latitude increases but increases after the latitude is greater than 20°. The decrease is caused by the general decrease of TC energy when it moves from south to north. The subsequent increase is possibly caused by the change in orientation of the coastal line in northern Mexico and southern Texas and higher mountains in northern Mexico, which leads to more chances of heavy TCP from landfalling storms. Part of this result agrees with what we have found in the elevation/TCP correlation for cluster 3 tracks. The event TCP has non-linearly responses to all first three location variables. The elevation variables (Figure 6d-j) are demonstrating mixed patterns. The TCP generally decreases as the mean elevation increases (Figure 6d) particularly from 0 to 1000 m, but it starts to increase when the elevation is greater than 2000 m. The maximum elevation has a similar pattern of change but the TCP increases with a larger magnitude at higher maximum elevations (> 2500 m). The TCP generally decreases monotonically with the minimum elevation (Figure 6f). The topography variables' influences on the TCP are more evident and consistent for range, standard deviation, and slope (Figure 6g, h, i). They are all showing a strong positive relationship with the TCP. All three variables describe different types of elevation variances within each 0.25° grid cell. Our RF models reflect that there is more TCP at places where the elevation is changing fast with large variance. The aspect of the slope (Figure 6j) is also demonstrating a nonlinear relationship with the TCP: the higher amount of TCP is observed for slopes that are facing the ocean (with aspect angle < 100° or > 250°, if we consider the profile of the coastline of Mexico) while less TCP is at the lee side slopes. In summary, the RF model well captures the combined and nonlinear influences from the locations and the topography to the ETCP variations. The pdps for the ETCP90 (Supplement 8) are showing similar patterns. The TCP show higher sensitivity to the longitude for more inland locations (< -100°). The range, standard deviation, and slope are all showing steeper curves within certain ranges (Supplement 8g, h, i). It indicates that the most extreme TCP events are possibly more sensitive to the topography changes, particularly where large local variations happen.

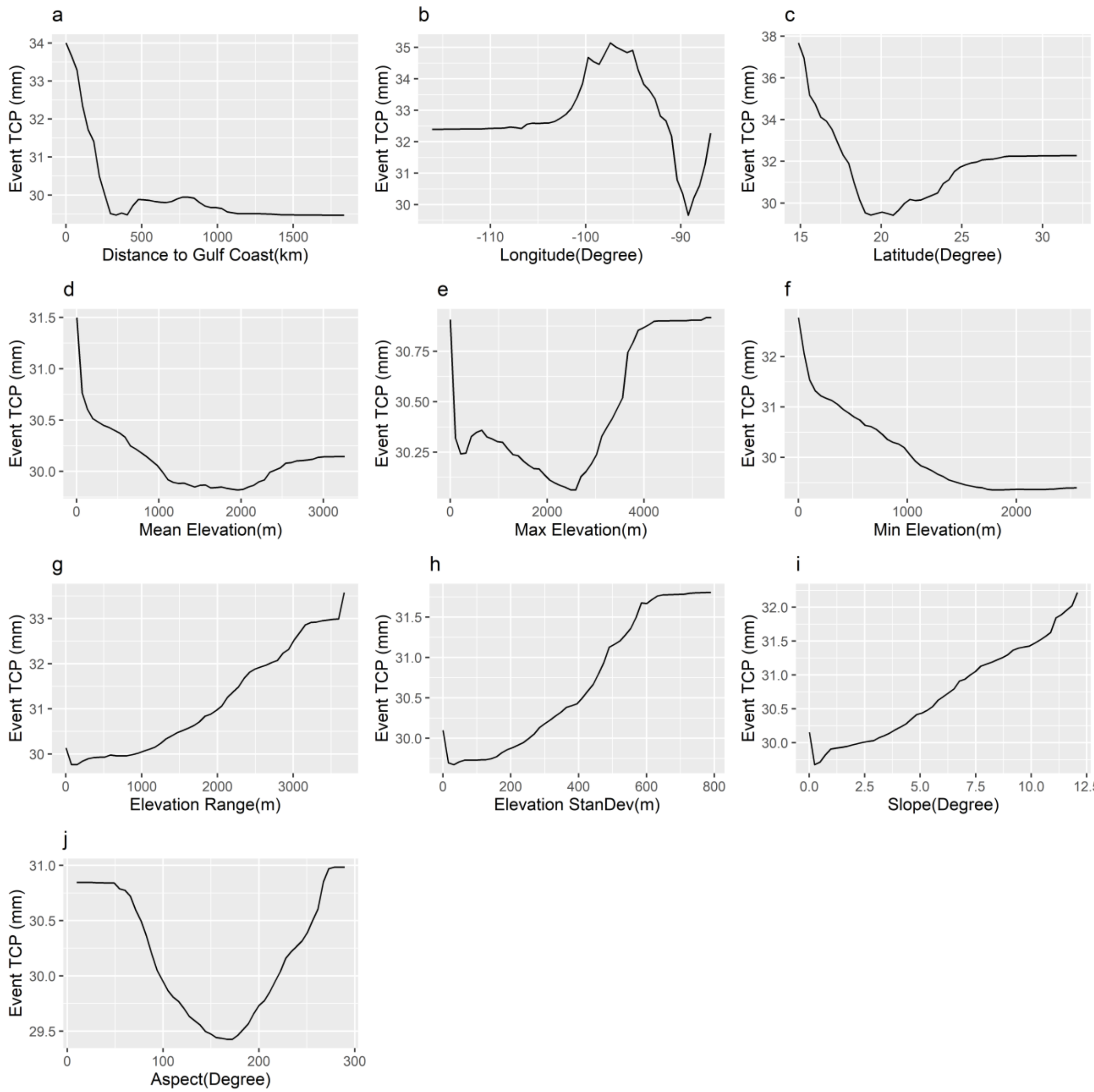


Figure 6. Partial Dependence Plot for static variables in the Whole Model for the ETCP

Pdps are demonstrating more variations for twelve dynamic variables (Figure 7 and Supplement 9). The distance to track is the most important variable in both ETCP and ETCP90 models. The TCP is very sensitive to its changes and the range is very large (~ 50 mm in Figure 7a and ~ 40 mm in Supplement 9a). The track cluster 3 storms produce the highest TCP, followed by track clusters 1 and 2 (Figure 7b). February to May have the highest event TCP while another peak happens between September and October (Figure 7c). Normally, the Atlantic hurricane season peaks in September, but it is also possible that the very rare storms not officially in the hurricane season have produced heavy precipitation and are reflected by the RF model. In the model for ETCP90 (Supplement 9c), October and November have the highest TCP. We have six variables representing the movement pattern of each storm. The forward U speed shows that more TCP is associated with storms with strong westward movement (Figure 7d). Storms with higher westward translation speed may have higher chances to make landfall in Mexico and the larger momentum to penetrate deeper inland and generate more precipitations. The TCP shows higher sensitivity to the forward V speed (30 mm in Figure 7e) than the U speed (10 mm in Figure 7d), which indicates that the north-south component of storm movement has a bigger impact on the event TCP than the east-west movement. Supplement 9e also shows that storms with a V speed between -5 to 5 knots are generating the most amount of extreme TCP. The forward Speed (Figure 7f) is a combination of both U Speed and V Speed and demonstrates more complex patterns. High TCP values are observed in storms moving below 5 knots but also in storms moving above 15 knots. The pdp plots of U, V, and mean forward speed for the ECTP90 (Supplement 9d, e, f) have similar patterns. The forward speed (Supplement 9f) shows a more consistent signal that more extreme TCP is associated with slow-moving storms (< 5 knots). The ETCP's response to the angle of the forward speed has two peaks at 305° and 320°

with a dip at ~ 310° (Figure 7h). The ETCP90 only has a higher value when the forward speed angle is between 290° to 310° (Supplement 9h). Those might be caused by the profile of the Mexico coastal line and the patterns in TC translation when they make landfall (e.g., angle to the coastlines when making landfall). Figures 7g and 7i show that more variances in the forward speed and its angle are likely to generate more TCP over the land. Variations in the storm tracks may be caused by TC's translations steered by the prevailing wind, the Beta effect, and interactions with other synoptic weather systems (Atallah et al., 2007) or track deflection from topography (Lin et al., 2002). Storms with complex tracks are reported to be big generators of the precipitation historically (e.g., Hurricane Harvey). It also shows that stalled storms generally make more precipitation than those not stalled (Figure 7k). Based on the annual TCP anomaly (Figure 2), the coastal grids (cluster 3) generally have a higher probability of receiving more ETCP than the inland grids (cluster 1 and 2) in Figure 7j. Finally, the Figure 7l confirms that the storms with longer durations are generating more TCP.

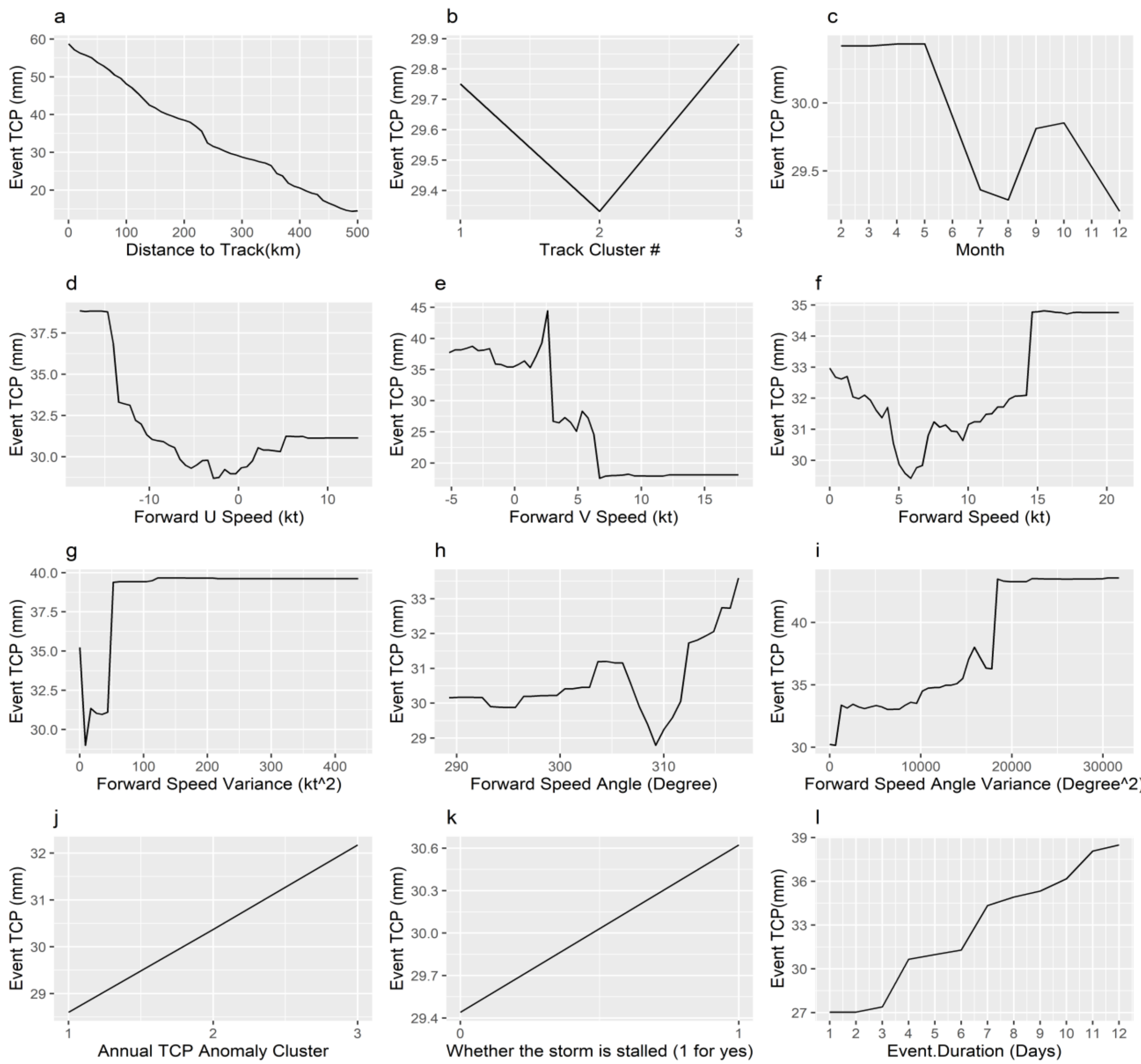


Figure 7. Partial Dependence Plot for Dynamic Variables in the Whole Model for the ETCP

3.4 **Extreme Cases**

Since the most extreme TCP events generated the largest damages, this section is focused on three storm events with the most extreme TCP in 99 years of climatology in Mexico. They are Hurricane Alex in 2010, Hurricane Igrid in 2013 and Major Hurricane Beulah in 1967. Alex and Igrid are originated from tropical disturbances from the Gulf of Mexico or the Caribbean Sea and experienced rapid intensification in a short translation distance before they made landfall. Beulah was originated from the Atlantic Ocean and gathered a large amount of energy through its long translation distance before it became the major hurricane that made landfall first in Texas. All three storms have produced > 400 mm precipitation at some locations (Figure 8a, d, and g) and those extreme precipitations caused massive flooding and landslides with losses of lives and infrastructures. The ETCP best model captures the spatial patterns of the TCP distributions very well for all three extreme cases (Figure 8b, e, h). Their scatter plots with the true observations agree very well with the y=x line and demonstrate high R^2 and low RMSE and MAE. The model still underpredicts > 300 mm TCP and they are mostly shown in Hurricane Alex and Igrid.

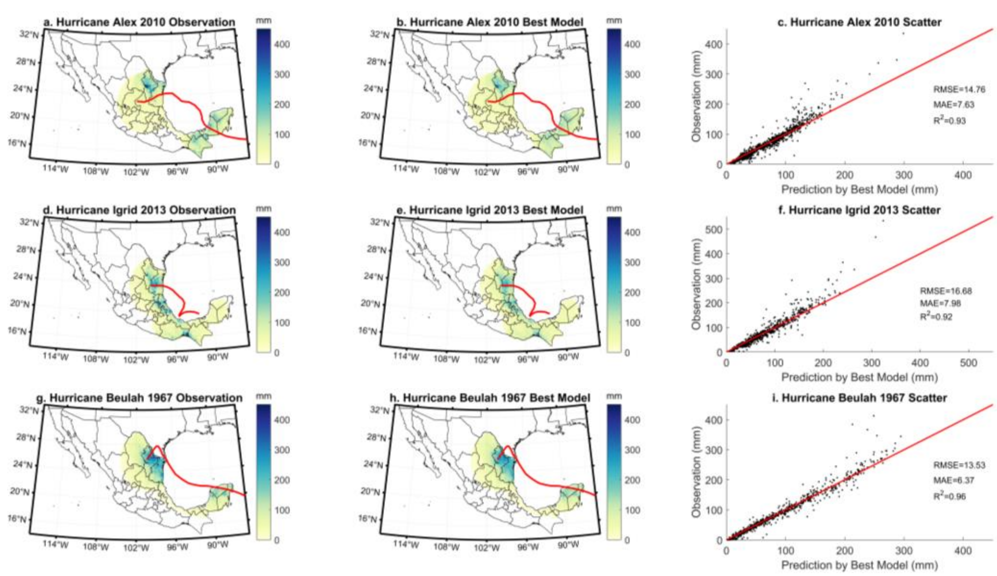


Figure 8. The precipitation of the three most intense TCP events from the observation and the Best ETCP Model

We also compared the extreme (> 90th percentile, P_{90}) TCP and the median range (between 45th percentile and 55th percentile) TCP samples and elevation variables associated with them. This comparison is finished for all three TCP anomaly grid clusters and all three storms. There are significant differences in the medians between the extreme and the median range TCP groups, ranging between 30 mm and 179 mm (Figure 9a, Supplement 10) with the maximized differences obtained by Hurricane Beulah. In most cases, the extreme TCP sample related elevation range and standard deviation have statistically significant larger median than those for the median range TCP sample (Figure 9b and c, Supplement 11 and 12, verified by Mann-Whitney Test at 95% level). This pattern is particularly stronger for cluster 1 and 2 locations, which are more inland and mountainous. In some cases, median range TCP samples have a larger elevation range and standard deviation than the extreme TCP samples. They are mostly happening in cluster 3 regions (coastal) in Hurricane Alex and Hurricane Beulah. The case study proves again that local topography variations have a strong enhancing effect for extreme TCP in Mexico, particularly over more inland regions.

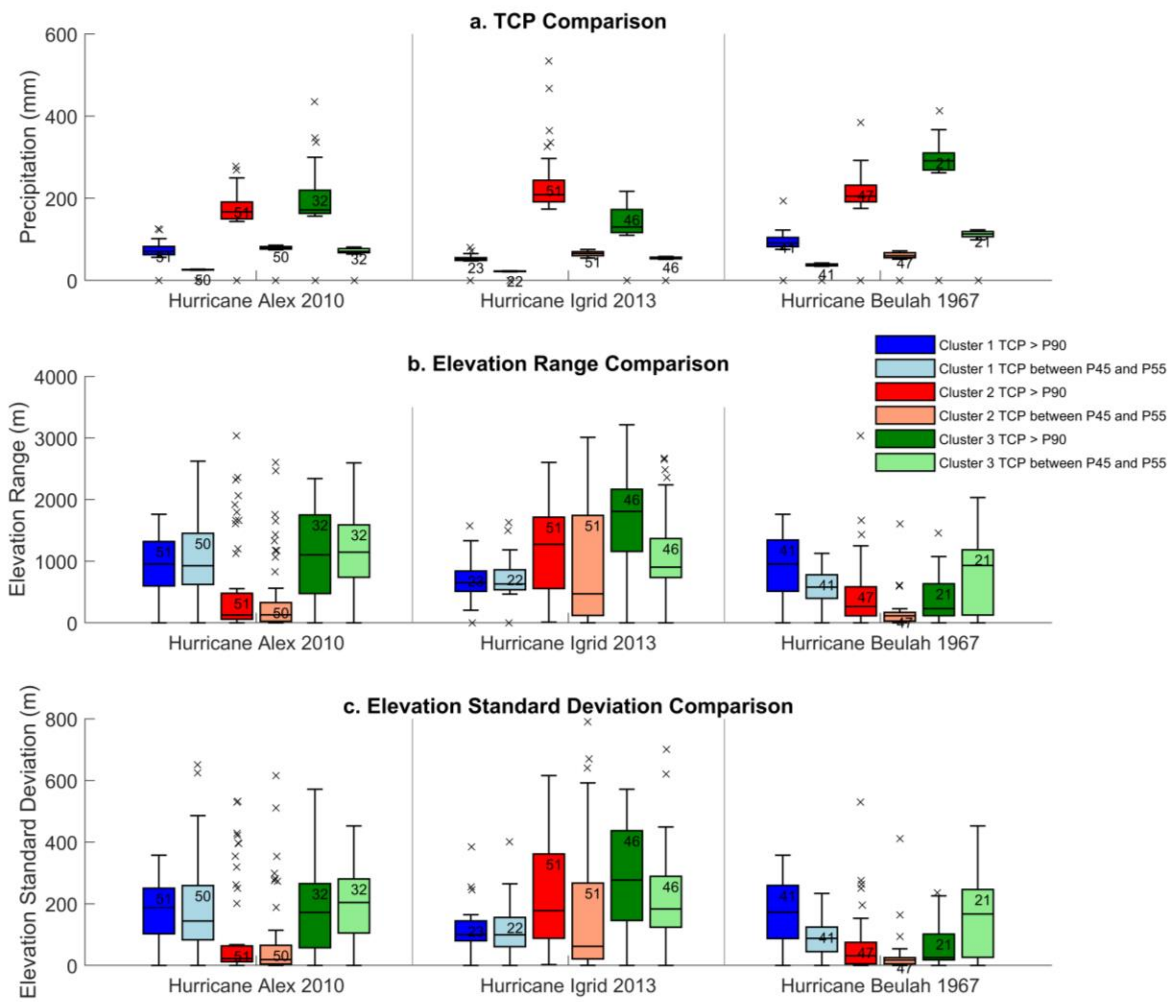


Figure 9. The comparison of topographic variables between locations with extreme TCP greater than the 90th Percentile (> P_{90}) and median range TCP (between P_{45} and P_{55}), separated by three annual TCP grid clusters.

4 Conclusion and Discussion

Many factors are influencing precipitation generated by TCs, which include their energy and moisture budget, storm size, and track characteristics, etc. Mexico is prone to strikes from heavy TCP events because of its long coastal lines and its complex terrain. However, how TCP changes spatially and temporally over Mexico and how different factors influence the overland TCP have not been thoroughly studied, particularly at the windward side of the Sierra Madre

Oriental. Our analysis is based on the longest available record from gauge observed daily TCP for Mexico since 1920 and we apply multiple data-mining approaches to understand this topic.

Strong decreasing gradients show in the annual mean TCP (AMTCP) and historical maximum event TCP (MAXETCP) from coast to inland. The clustered correlation analysis demonstrates that location variables have the most consistent and strongest correlations with the AMTCP and MAXETCP. Elevation variables show mixed correlations with the TCP, diversified by locations and elevation variable types. The elevation range, standard deviation and slope show positive correlations with the TCP, particularly for inland areas, while the mean, max and min elevations show more negative correlations for coastal areas. The reason is that the elevations are also highly correlated with their locations in Mexico. The clustered correlation have filtered out some impacts from the locations to elevation's impact to TCP but are not able to completely filter them out. Indeed, locations' influences on AMTCP and MAXETCP are so strong that the best RF models only choose three location variables (latitudes, longitude, and distance to the coast) and can explain most of the variance in AMTCP and MAXETCP with very little cross-validation error.

While three location variables can explain most of the variance in AMTCP and MaxETCP, we have more variables (both static and dynamic) to model the much more complex variations in event TCP. Although there are high diversity and complexity in the variables used by the best models for the ETCP (18 variables) and ETCP90 (17 variables), most of the relationships with the TCP can be explained well by their VI and partial dependence plots. Many variables show a similar pattern of influences to TCP as demonstrated by the correlation analysis, but with additional details and non-linear relationships. We find that the distance to the track is the most important factor that determines the event TCP in our model. It ranks highest in variable importance and the event TCP has a very high sensitivity to it. Longitude, latitude, and distance to the coast are the three most important static variables in the model. There is a strong decreasing gradient in the possibility of TCP from the coastal area to inland, and the TCP probability is changing with latitude and longitude, controlled by both the decaying of the TC energy, the profile of the coastal line, and the moving direction of the TC. The translation characteristics of the storm are another group of dynamic variables that are important to the event TCP variations. Slower moving storms (particularly in the north-south direction) are generally producing heavier event TCP because there is a longer duration of the storm at a specific location. Many slower-moving storms have generated the worst inland flooding event and Kossin (2018) shows that the TCs were moving slower globally in recent years and possibly generated more precipitation. Our model also shows that more variations in the storm moving speed and angle are contributing more event TCP and stalled TCs are also likely to generate more TCP. Stalled storms are special cases and are sometimes particularly dangerous because the convection is lifted suddenly by other synoptic systems, which speeds up the condensation of water vapor. And they may also stay longer with their bent tracks and generate more precipitation. Hall and Kossin (2019) also demonstrate that the Atlantic TCs have been stalled more frequently in recent years, which may introduce more probability of extreme precipitation events with long duration like Hurricane Harvey. Finally, the topographic variables also play important roles in our RF models, particularly for extreme cases. We show nonlinear relationships between elevation variables and the TCP in our models. Higher TCP cases are most likely located at coastal areas with lower mean elevation, while regions with higher elevation are also likely to have less frequent but very high TCP events. The range, standard deviation and slope are demonstrating a monotonically enhancing relationship with the TCP. This relationship demonstrates both in the correlation and the RF analyses but particularly stronger over more inland areas. Lastly, more windward slopes have higher TCP than leeward ones.

The RF model is an effective machine learning tool to explore important factors that influence the TCP overland and their complex relationships in the process. Our model results at both annual and event scale demonstrate that the RF model excels in the fitting and prediction skills than traditional statistical models. Our best RF models obtain 95% explained variances of the Event TCP (ETCP) and 98% explained variance of the AMTCP, both estimated from multiple cross-validations. They have significantly improved the previously reported performance of the linear regression model for the annual precipitation in different mountainous areas (31 to 75% variance explained) around the world (Basist et al., 1994). The ETCP model shows excellent error statistics (MAE and RMSE) when making out of sample predictions, and the ETCP90 model improves the prediction skills of the ETCP model for the extreme TCPs. The ETCP model can also predict extreme event TCP cases with good agreement to the observed spatial patterns.

Our study shows a promising future for the application of this type of machine learning technique in operational TCP forecasting, which relies on the accuracy of ensemble TC track forecasting and other available information as inputs. The execution of our current RF model is very efficient so it can give skillful predictions of the TCP with a short preparation and waiting time, which provides valuable preparation and response time for incoming extreme TCP related disasters. Our current study looks at factors including locations, topography, storm tracks, storm translation pattern, storm duration, etc. We believe that there are many more dynamic factors contributing to the TCP variations at different scales, which may include the sea surface temperature, the El Niño–Southern Oscillation (ENSO), energy and moisture budget over the land, vertical wind shear, extratropical transition (ET) of the TC, and TC's interactions with other synoptic systems. It will be interesting to develop machine learning models at other temporal scales (annual, daily, or hourly) using other independent precipitation datasets. The current RF model still needs improvements in skills of predicting the most extreme TCP cases.

Acknowledgments

This research is supported by the NSF Grant #1619681: The Michigan Louis Stokes Alliance for Minority Participation (MI-LSAMP)

We thank Dr. Steven Quiring from the Ohio State University for offering the HPC access to the Ohio Supercomputer Center (OSC) for the model development and cross-validation. We also thank the three anonymous reviewers for their substantial contributions in improving this work.

Gauge observations are derived from the Daily Global Historical Climatology Network (GHCN-D) (https://www.ncdc.noaa.gov/ghcnd-data-access) and obtained from the National Weather Service of Mexico by requiring. Tropical cyclone tracks are obtained from International Best Track Archive for Climate Stewardship (IBTrACS) (https://www.ncdc.noaa.gov/ibtracs/). The DEM data is obtained from the USGS EROS Archive - Digital Elevation - Global 30 Arc-Second Elevation (GTOPO30) (https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30?qt-science_center_objects=0#qt-science_center_objects)

References

Agustín Breña-Naranjo, J., Pedrozo-Acuña, A., Pozos-Estrada, O., Jiménez-López, S. A., & López-López, M. R. (2015). The contribution of tropical cyclones to rainfall in Mexico. *Physics and Chemistry of the Earth, Parts A/B/C, 83-84*, 111-122. doi:https://doi.org/10.1016/j.pce.2015.05.011

Arndt, D. S., Basara, J. B., McPherson, R. A., Illston, B. G., McManus, G. D., & Demko, D. B. (2009). Observations of the Overland Reintensification of Tropical Storm Erin (2007). *Bulletin of the American Meteorological Society, 90*(8), 1079-1094. doi:10.1175/2009bams2644.1

Atallah, E., Bosart, L. F., & Aiyyer, A. R. (2007). Precipitation Distribution Associated with Landfalling Tropical Cyclones over the Eastern United States. *Monthly Weather Review, 135*(6), 2185-2206. doi:10.1175/mwr3382.1

Best, D. J., & Roberts, D. E. (1975). Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 24*(3), 377-379. doi:10.2307/2347111

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/a:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. U.S.A: Taylor & Francis.

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of geographical information systems*. Oxford: Oxford University Press.

Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P., & Ghil, M. (2007). Cluster Analysis of Typhoon Tracks. Part I: General Properties. *Journal of Climate, 20*(14), 3635-3653. doi:10.1175/jcli4188.1

Chan, K. T. F. (2019). Are global tropical cyclones moving slower in a warming climate? *Environmental Research Letters, 14*(10), 104015. doi:10.1088/1748-9326/ab4031

Emanuel, K. (2017). Assessing the present and future probability of Hurricane Harvey’s rainfall. *Proceedings of the National Academy of Sciences, 114*(48), 12681-12684. doi:10.1073/pnas.1716222114

Farfán, L. M., & Cortez, M. (2005). An Observational and Modeling Analysis of the Landfall of Hurricane Marty (2003) in Baja California, Mexico. *Monthly Weather Review, 133*(7), 2069-2090. doi:10.1175/mwr2966.1

Farfán, L. M., & Zehnder, J. A. (2001). An Analysis of the Landfall of Hurricane Nora (1997). *Monthly Weather Review, 129*(8), 2073-2088. doi:10.1175/1520-0493(2001)129<2073:Aaotlo>2.0.Co;2

Franco-Díaz, A., Klingaman, N. P., Vidale, P. L., Guo, L., & Demory, M.-E. (2019). The contribution of tropical cyclones to the atmospheric branch of Middle America's hydrological cycle using observed and reanalysis tracks. *Climate Dynamics, 53*(9-10), 6145-6158. doi:10.1007/s00382-019-04920-z

Gaffney, S. J., Robertson, A. W., Smyth, P., Camargo, S. J., & Ghil, M. (2007). Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dynamics, 29*(4), 423-440. doi:10.1007/s00382-007-0235-z

Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal, 9*(1), 421–436.

Hall, T. M., & Kossin, J. P. (2019). Hurricane stalling along the North American coast and implications for rainfall. *npj Climate and Atmospheric Science, 2*(1). doi:10.1038/s41612-019-0074-8

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition*: New York: Springer.

Houze, R. A. (2012). Orographic effects on precipitating clouds. *Reviews of Geophysics, 50*(1). doi:10.1029/2011rg000365

Huang, C., Chou, C., Chen, S., & Xie, J. (2020). Topographic Rainfall of Tropical Cyclones past a Mountain Range as Categorized by Idealized Simulations. *Weather and Forecasting, 35*(1), 25-49. doi:10.1175/waf-d-19-0120.1

Huang, J.-C., Yu, C.-K., Lee, J.-Y., Cheng, L.-W., Lee, T.-Y., & Kao, S.-J. (2012). Linking typhoon tracks and spatial rainfall patterns for improving flood lead time predictions over a mesoscale mountainous watershed. *Water Resources Research, 48*(9), n/a-n/a. doi:10.1029/2011wr011508

Kepert, J. (2001). The Dynamics of Boundary Layer Jets within the Tropical Cyclone Core. Part I: Linear Theory. *Journal of the Atmospheric Sciences, 58*(17), 2469-2484. doi:10.1175/1520-0469(2001)058<2469:Tdoblj>2.0.Co;2

Kimball, S. K. (2008). Structure and Evolution of Rainfall in Numerically Simulated Landfalling Hurricanes. *136*(10), 3822-3847. doi:10.1175/2008mwr2304.1

Knutson, T., Camargo, S. J., Chan, J. C. L., Emanuel, K., Ho, C.-H., Kossin, J., . . . Wu, L. (2019). Tropical Cyclones and Climate Change Assessment: Part I: Detection and Attribution. *Bulletin of the American Meteorological Society, 100*(10), 1987-2007. doi:10.1175/bams-d-18-0189.1

Knutson, T., Camargo, S. J., Chan, J. C. L., Emanuel, K., Ho, C.-H., Kossin, J., . . . Wu, L. (2020). Tropical Cyclones and Climate Change Assessment: Part II: Projected Response to Anthropogenic Warming. *Bulletin of the American Meteorological Society, 101*(3), E303-E322. doi:10.1175/bams-d-18-0194.1

Kossin, J. P. (2018). A global slowdown of tropical-cyclone translation speed. *Nature, 558*(7708), 104-107. doi:10.1038/s41586-018-0158-3

Langousis, A., & Veneziano, D. (2009). Theoretical model of rainfall in tropical cyclones for the assessment of long-term risk. *Journal of Geophysical Research, 114*(D2). doi:10.1029/2008jd010080

Li, Y., Huang, W., & Zhao, J. (2007). Roles of mesoscale terrain and latent heat release in typhoon precipitation: A numerical case study. *Advances in Atmospheric Sciences, 24*(1), 35-43. doi:10.1007/s00376-007-0035-8

Lin, Y.-L., Chen, S.-Y., Hill, C. M., & Huang, C.-Y. (2005). Control Parameters for the Influence of a Mesoscale Mountain Range on Cyclone Track Continuity and Deflection. *Journal of the Atmospheric Sciences, 62*(6), 1849-1866. doi:10.1175/jas3439.1

Lin, Y.-L., Ensley, D. B., Chiao, S., & Huang, C.-Y. (2002). Orographic Influences on Rainfall and Track Deflection Associated with the Passage of a Tropical Cyclone. *Monthly Weather Review, 130*(12), 2929-2950. doi:10.1175/1520-0493(2002)130<2929:Oiorat>2.0.Co;2

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist., 18*(1), 50-60. doi:10.1214/aoms/1177730491

Mascaro, G., Vivoni, E. R., Gochis, D. J., Watts, C. J., & Rodriguez, J. C. (2014). Temporal Downscaling and Statistical Analysis of Rainfall across a Topographic Transect in Northwest Mexico. *Journal of Applied Meteorology and Climatology, 53*(4), 910-927. doi:10.1175/jamc-d-13-0330.1

Matyas, C. (2007). Quantifying the Shapes of U.S. Landfalling Tropical Cyclone Rain Shields*. *The Professional Geographer, 59*(2), 158-172. doi:10.1111/j.1467-9272.2007.00604.x

McRoberts, D. B., Quiring, S. M., & Guikema, S. D. (2018). Improving Hurricane Power Outage Prediction Models Through the Inclusion of Local Environmental Factors. *Risk Analysis, 38*(12), 2722-2737. doi:10.1111/risa.12728

Nateghi, R., Guikema, S., & Quiring, S. M. (2014). Power Outage Estimation for Tropical Cyclones: Improved Accuracy with Simpler Models. *Risk Analysis, 34*(6), 1069-1078. doi:10.1111/risa.12131

Pineda-Martinez, L. F., & Carbajal, N. (2009). Mesoscale numerical modeling of meteorological events in a strong topographic gradient in the northeastern part of Mexico. *Climate Dynamics, 33*(2-3), 297-312. doi:10.1007/s00382-009-0549-0

Ramsay, H. A., & Leslie, L. M. (2008). The Effects of Complex Terrain on Severe Landfalling Tropical Cyclone Larry (2006) over Northeast Australia. *Monthly Weather Review, 136*(11), 4334-4354. doi:10.1175/2008mwr2429.1

Risser, M. D., & Wehner, M. F. (2017). Attributable Human-Induced Changes in the Likelihood and Magnitude of the Observed Extreme Precipitation during Hurricane Harvey. *Geophysical Research Letters, 44*(24), 12,457-412,464. doi:10.1002/2017gl075888

Shapiro, L. J. (1983). The Asymmetric Boundary layer Flow Under a Translating Hurricane. *Journal of the Atmospheric Sciences, 40*(8), 1984-1998. doi:10.1175/1520-0469(1983)040<1984:Tablfu>2.0.Co;2

Skok, G., Bacmeister, J., & Tribbia, J. (2013). Analysis of Tropical Cyclone Precipitation Using an Object-Based Algorithm. *Journal of Climate, 26*(8), 2563-2579. doi:10.1175/jcli-d-12-00135.1

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. doi:10.1186/1471-2105-8-25

Trenberth, K. E., Cheng, L., Jacobs, P., Zhang, Y., & Fasullo, J. (2018). Hurricane Harvey Links to Ocean Heat Content and Climate Change Adaptation. *Earth's Future, 6*(5), 730-744. doi:10.1029/2018ef000825

Tuleya, R. E. (1994). Tropical Storm Development and Decay: Sensitivity to Surface Boundary Conditions. *Monthly Weather Review, 122*(2), 291-304. doi:10.1175/1520-0493(1994)122<0291:tsdads>2.0.co;2

Vivoni, E. R., Gutiérrez-Jurado, H. A., Aragón, C. A., Méndez-Barroso, L. A., Rinehart, A. J., Wyckoff, R. L., . . . Jackson, T. J. (2007). Variation of Hydrometeorological Conditions along a Topographic Transect in Northwestern Mexico during the North American Monsoon. *Journal of Climate, 20*(9), 1792-1809. doi:10.1175/jcli4094.1

Wu, C. C., Yen, T. H., Kuo, Y. H., & Wang, W. (2002). Rainfall simulation associated with typhoon herb (1996) near Taiwan. Part I: The topographic effect. *Weather and Forecasting, 17*(5), 1001-1015. doi:10.1175/1520-0434(2003)017<1001:Rsawth>2.0.Co;2

Zehnder, J. A. (1993). The Influence of Large-Scale Topography on Barotropic Vortex Motion. *Journal of the Atmospheric Sciences, 50*(15), 2519-2532. doi:10.1175/1520-0469(1993)050<2519:Tiolst>2.0.Co;2

Zhang, W., Villarini, G., Vecchi, G. A., & Smith, J. A. (2018). Urbanization exacerbated the rainfall and flooding caused by hurricane Harvey in Houston. *Nature, 563*(7731), 384-388. doi:10.1038/s41586-018-0676-z

Zhou, Y., Matyas, C., Li, H., & Tang, J. (2018). Conditions associated with rain field size for tropical cyclones landfalling over the Eastern United States. *Atmospheric Research, 214*, 375-385. doi:10.1016/j.atmosres.2018.08.019

Zhu, L., & Quiring, S. M. (2013). Variations in tropical cyclone precipitation in Texas (1950 to 2009). *Journal of Geophysical Research: Atmospheres, 118*(8), 3085-3096. doi:10.1029/2012jd018554

Zhu, L., & Quiring, S. M. (2017). An Extraction Method for Long-Term Tropical Cyclone Precipitation from Daily Rain Gauges. *Journal of Hydrometeorology, 18*(9), 2559-2576. doi:10.1175/jhm-d-16-0291.1