Applying a Bayesian network based on Gaussian copulas to model the hydraulic boundary conditions for hurricane flood risk analysis in a coastal watershed

Sebastian, Antonia; Dupuits, Guy; Morales Napoles, Oswaldo

Publication date
2017

Document Version
Accepted author manuscript

Published in
Coastal Engineering

Citation (APA)
Sebastian, A., Dupuits, G., & Morales Napoles, O. (2017). Applying a Bayesian network based on Gaussian copulas to model the hydraulic boundary conditions for hurricane flood risk analysis in a coastal watershed. *Coastal Engineering*, *125*, 42-50.

Postprint of Coastal Engineering
Volume 125, July 2017, Pages 42–50
Link to formal publication (Elsevier): http://dx.doi.org/10.1016/j.coastaleng.2017.03.008

Applying a Bayesian Network based on Gaussian Copulas to Model the Hydraulic Boundary Conditions for Hurricane Flood Risk Analysis in a Coastal Watershed

A. Sebastian[a,b,*], E.J.C. Dupuits[b], O. Morales-Nápoles[b]

[a] *Department of Civil and Environmental Engineering, Rice University, Houston, Texas*
[b] *Faculty of Civil Engineering and Geosciences, Delft Technical University, Delft, The Netherlands*

Abstract

In recent years significant emphasis has been placed on quantifying coastal flood hazards in the U.S. using high resolution 2-D hydrodynamic and nearshore wave models. However, these studies are computationally expensive and often neglect to consider the flooding that arises from the combined hazards of precipitation and storm surge in coastal watersheds. This paper describes a method to stochastically simulate a large number of combinations of peak storm surge and cumulative precipitation to determine the hydraulic boundary conditions for a low-lying coastal watershed draining into a semi-enclosed tidal bay. The method is computationally efficient and takes into consideration five tropical cyclone characteristics at landfall: windspeed, angle of approach, landfall location, radius of maximum winds, and forward speed. A precipitation gage network and tidal gage data were used, along with observations from over 300 tropical cyclones in the Gulf of Mexico. A Non-parametric Bayesian Network was built to generate 100,000 synthetic storm events and used as input to an empirical wind set-up model to simulate storm surge within a tidal bay and at the downstream boundary of the watershed. Based on the results, probable combinations of cumulative precipitation and peak storm surge for the watershed during hurricane conditions are determined. These boundary conditions can be easily incorporated into a coastal riverine model to determine flood risk in the watershed.

*Corresponding author
Email address: a.g.sebastian@tudelft.nl (A. Sebastian)

Keywords: tropical cyclone, flood risk, compound flooding, storm surge, Bayesian networks, copula
0000 MSC: 00-01, 99-00

1. Introduction

Integrated flood risk assessment is critical for the determination of appropriate flood mitigation strategies for heavily populated, low-lying coastal areas. While the number of deaths from tropical cyclones have decreased with the development of advanced prediction and early warning systems, economic losses have increased exponentially due to rapid population growth and urban development near the coast. It is estimated that today almost half of the global population lives within 150 km of a coastline [1]. These highly urbanized coastal areas are threatened by the combined impacts of severe storms, especially hurricane-induced storm surge and heavy precipitation. Communities along the U.S. Gulf Coast and in delta regions around the world, where storm surge often coincides with heavy precipitation, are especially vulnerable.

In the United States, the 100-year Federal Emergency Management Agency (FEMA) floodplain is used as the primary instrument for delineating and mitigating flood risk. This boundary, indicating the 1% percent chance of inundation each year from riverine *or* coastal flooding, drives federal flood insurance requirements, household protective actions, and local mitigation policies. It also determines where future development can take place and what it will look like. However, increasing evidence suggests that the FEMA 100-year floodplain is a poor predictor of actual flood damage and that in some watersheds, upwards of 50% of insured losses are occurring outside of the demarcated flood hazard areas [2]. This is especially apparent along the Gulf Coast where insured assets account for 41% of flood insurance policies in the United States, but amounted to more than 80% of claim payouts between 1978 and 2010 [3].

One of the primary contributing factors to floodplain inaccuracy is the age of existing FEMA floodplain maps; more than 60% of the maps are at least 10 years old and in some coastal areas, maps date back as far as the mid-to-late 1970s [4, 5]. In addition to their age, riverine floodplains are derived using deterministic hydrologic and hydraulic models based on a single design storm. This leads to

compounding uncertainties since the modeled floodplain is only as accurate as the information used to define them, including the original assumptions made about the design storm. Other sources of inaccuracy include: limited historical hydrometeorological observations [6]; spatio-temporal variations in precipitation [7]; and changes in climate, topography, and land use conditions [8].

Similarly, early FEMA coastal floodplains were generated using a design storm, or Standard Project Hurricane (SPH), which was intended to represent a probable, yet infrequent hurricane along a section of the coast [9]. These design storms were often based on a single storm characteristic: intensity, derived from historical storm data prior to 1960. The relative calm in the period prior to 1960 and the oversimplification of hurricane behavior typically led to an underestimation of surge heights at a particular coastal location [10]. In subsequent studies, attempts were made to assess return period surge using historical water level records; however, the lack of extreme water level events at a single coastal location coupled with anonymously high single-storm records made these estimates difficult to resolve [10]. In response, a move toward generating a probabilistic suite of storms was made in the 1970s and 1980s and a Joint Probability Method (JPM) approach based on cluster analysis of hurricane characteristics at landfall was developed to generate parametric wind fields at landfall [11]. However, the sources of uncertainty in this approach were quite high, primarily due to lack of historical data [10].

With the advent of complex computing and high-resolution storm surge and wave models, renewed interest in improving joint probability methods (JPM) for coastal modeling has taken place. Recently, JPM-Optimum Sampling (JPM-OS) was introduced as a method to reduce the number of representative synthetic storms needed for simulating storm surge at any particular section of the coast. This method is being applied in many of the current FEMA Flood Insurance Studies (to be completed in 2020). However, even with the application of the JPM-OS approach, probabilistic modeling using high-resolution storm surge models is computationally demanding, requiring large computer clusters to run [10, 12, 13]. Furthermore, despite the marked improvement in coastal modeling and storm surge mapping, the new FEMA floodplain maps still neglect to consider the flood risk resulting from the interaction between rainfall-runoff

and storm surge at the coast. In small, low-lying coastal watersheds, where hydrologic response to precipitation is nearly instantaneous, determining the joint exceedance of precipitation and storm surge is critical for assessing flood risk. Here, small variations in downstream elevation extend far upstream, impeding the propagation and exit of the precipitation-induced flood wave from the watershed.

This paper introduces a method for quantifying hurricane boundary conditions for small, urbanized coastal watersheds draining into a tidally influenced, semi-enclosed bay systems using a non-parametric Bayesian network (NPBN) based on Gaussian copulas. The NPBN is used to generate a suite of synthetic tropical cyclones and the events are input into an empirical wind setup model to stochastically simulate a large number of storms in the bay system. The modeled combinations of storm surge and precipitation provide an initial estimate of joint exceedance probabilities for a coastal watershed.

As a case study, the method is applied to the Clear Creek Watershed located 32 km southeast of Houston, Texas on the west side of Galveston Bay (see Section 3). Galveston Bay creates a complex environment where surge is often higher on the north and west side of the bay than on the east due to local wind-setup caused by counter-clockwise hurricane winds over the Bay [14]. Furthermore, the combined impacts of little topographic relief, slow or limited infiltration, rapid urban development, regional subsidence and sea level rise, and intense storms have led to frequent and severe flooding in the watershed. In the next section, we provide an overview of non-parametric Bayesian networks and introduce key terms. This is followed by a description of the study area and an overview of the method, model results, discussion and conclusion.

2. Bayesian Networks

Bayesian Networks (BN) are probabilistic graphical models that can be used to represent a large number of interdependent variables [15, 16, 17]. The variables in the network can be either discrete or continuous, and their dependency on one another is quantified by conditional probability functions. One of the primary advantages of BNs is that the probability distribution functions in the graph can be easily updated to reflect changes in the joint distribution (i.e., in-

ference). Given their characteristics, BNs can be efficiently sampled to generate large synthetic data sets [18].

BNs are composed of a number of children (successor) and parent (predecessor) "nodes", which represent a set of random variables $(X_1, X_2, ..., X_n)$. In this paper, the nodes are labeled on the set of positive integers. An ordered pair of elements of this set is called an "arc" which represents the dependence between each parent-child pair in the graphical network and has a defined direction such that the graph remains acyclic. Together, the nodes and arcs represent the dependence structure between the variables in the model, where the joint distribution of the child nodes can be computed as a product of conditional probability functions:

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i | \mathbf{x}_{Pa(X_i)}) \tag{1}$$

where $Pa(X_i)$ is the set of parent nodes of x_i, with $i = 1, ..., n$. For nodes without parents, $Pa(X_i) = \emptyset$ and the marginal density is used in Equation 1.

In this study, a non-parametric continuous Bayesian network (NPBN) based on copulas is applied as presented in [17, 19, 18]. In this model, each continuous random variable is represented by its empirical distribution (hence the non-parametric part of the name in this class of BNs) and the dependence structures in the network are built using bivariate copula of the one-parameter class. Although the term semi-parametric BNs may be more appropriate to describe this type of BN (since the procedure relies on one-parameter copulas), to remain consistent with the previous literature, we use the term NPBN in this paper. An example NPBN with three nodes is shown in Figure 1.

NPBNs have several advantages over traditional regression methods for application to environmental data. For example, environmental data often exhibits non-linear behavior making it difficult to represent using traditional regression models. NPBNs can capture and model the interdependencies between complex environmental variables. Moreover, the graphical nature of a NPBN makes the dependence configuration between environmental variables explicit [20].

In an NPBN, the arcs between each parent-child pair (i.e., $Pa(X_i) \rightarrow X_i$) are represented by one-parameter (conditional) copulas. In its most general form, the bivariate cumulative distribution function $F_{X_i,X_j}(x_i, x_j)$ of the ran-

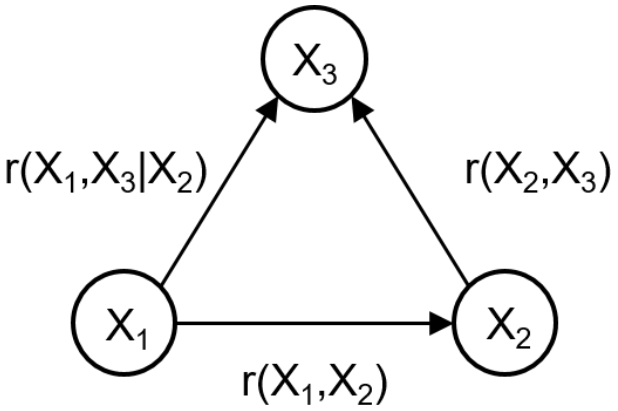

Figure 1: Example non-parametric Bayesian network (NPBN) with three nodes (X_1,...X_3) connected by three arcs. In this paper, each arc represents a bivariate (conditional) copula of the one-parameter class.

dom variables X_i and X_j becomes

$$F_{X_i,X_j}(x_i, x_j) = C[F_{X_i}(x_i), F_{X_j}(x_j)] \tag{2}$$

where $F_{X_i}(x_i)$ and $F_{X_j}(x_j)$ are the marginal distributions of X_i and X_j and there exists a copula, C, that describes their dependence structure [21]. The advantage of using copulas is that the dependence structure of $F_{X_i,X_j}(x_i, x_j)$ is captured by the copula and the selection of the copula is independent of the marginal distributions of X_i and X_j [22, 23, 24].

To build a continuous NPBN, it is necessary to specify an empirical marginal distribution for each variable and a rank correlation coefficient for each arc. Like bivariate copulas, which are parameterized by (conditional) product moment correlations, the NPBN is parameterized using (conditional) rank correlations. The rank correlation measures the monotonic dependence between the cumulative distribution functions of two random variables. The rank correlation, r, of two variables, X_i and X_j, is

$$r(X_i, X_j) = \rho(F_{X_i}(x_i), F_{X_j}(x_j)) \tag{3}$$

where ρ is the product moment correlation of the cumulative distribution functions F_{X_i} and F_{X_j} of X_i and X_j respectively. There exists a non-unique structure of the NPBN such that the (conditional) rank correlation for each variable,

X_i, with m parents, $Pa_1(X_i), ..., Pa_m(X_i)$, the arc, $Pa_j(X_i) \rightarrow X_i$, can be written as

$$\begin{cases} r(X_i, Pa_j(X_i)), j = 1 \\ r(X_i, Pa_j(X_i)|Pa_1(X_i), ..., Pa_{j-1}(X_i)), j = 2, ..., m \end{cases} \tag{4}$$

where j is the non-unique sampling order.

The assignment of the sampling order in Equation 4 in the case of the Gaussian copula necessitates that the resulting correlation matrix is positive definite. However, depending on the assignment and ordering of arcs of the NPBN, the unconditional rank correlations will be calculated recursively. For example, suppose the product moment correlation matrix of interest has elements $\rho_{1,2}$, $\rho_{1,3}$ and $\rho_{2,3}$, where $\rho_{1,3}$ is related to $\rho_{1,3|2}$ (which was specified in the assignment (4) through $r_{1,3|2}$, together with $r_{1,2}$ and $r_{2,3}$) through the expression

$$\rho_{1,3|2} = \frac{\rho_{1,3} - \rho_{1,2} \cdot \rho_{2,3}}{\sqrt{(1 - \rho_{1,2}^2)(1 - \rho_{2,3}^2)}}. \tag{5}$$

and $r_{i,j} = (\frac{6}{\pi}) \cdot \arcsin(\frac{\rho_{i,j}}{2})$. This structure is shown in Figure 1.

Notice that for the small example above, an alternative assignment could have been $r_{1,2}$, $r_{1,3}$ and $r_{2,3|1}$. For these two small examples, the resulting dependence structure of the NPBN through the Gaussian copula would be identical; however, it becomes clear that for larger structures this would not be the case. The dependence structure (summarized as a correlation matrix) can be further investigated using the validation procedures described in Appendix A.3. For a complete overview of NPBNs the reader is referred to [18].

Prior to 2003, the application of copulas to natural systems was limited, resulting in only a handful of relevant publications between 1979 and 2003 [25]. Since then, the volume of literature in the hydrological sciences has increased significantly. Copulas have been widely applied to hydrologic problems ranging from the analysis of return period flows in river systems [26, 20, 27] to precipitation forecasts [28] and, more recently, water levels in coastal systems [29, 30, 31, 32]. Copulas have also been applied to study hurricane characteristics in the U.S., but have been limited to analyzing the joint probabilities of wind and surge [33]. To our knowledge, this paper provides the first example of an extension of bivariate copulas to multivariate prediction of hurricane characteristics for the Gulf of Mexico.

3. Study Area

The Houston-Galveston region lies on the banks of Galveston Bay and is the fastest growing coastal area in the Gulf of Mexico. Galveston Bay is the seventh largest estuary in the United States. It has a surface area of approximately 1554 km^2 and is a shallow, wind-driven system with an average depth of approximately 3 meters. The Bay is separated from the Gulf of Mexico by two barrier islands: Bolivar and Galveston. Currently, more than 1.6 million people live in the Hurricane Evacuation Zones bordering Galveston Bay and it is projected that this number will approach 2.4 million by 2035 [34].

On average, the Houston-Galveston region experiences a tropical cyclone once every nine years and a major hurricane (Category 3 or greater) once every 25 years [35]. During the 1900 Hurricane, surge at the coast exceeded 4 m and to date, it remains the deadliest hurricane in U.S. history. While tropical cyclones have the potential to cause considerable storm surge in the region, they also bring with them heavy rainfall. The most notable rainfall events include Tropical Storm Claudette (1979) and Tropical Storm Allison (2001). Wahl et al. [36] note that there is a high risk of compound flooding in the Houston-Galveston region and that the frequency with which compound flood events occur along the U.S. Atlantic and Gulf Coasts is expected to increase under changing climate conditions.

In this paper we focus on the Clear Creek Watershed, located in the rapidly developing region on the west side of Galveston Bay, 32 km south of Houston (Figure 2). The Watershed is approximately 72 km long and covers 666 km^2. It is drained by two primary tributaries that feed Clear Lake, an estuarine lake with an average depth of 1.0 m, which empties into Galveston Bay via two outlets near Kemah, Texas. This analysis focuses on determining boundary conditions (i.e., storm surge and precipitation) for the Clear Creek Watershed. Like at many locations along the Gulf of Mexico, historical data near the watershed outlet is limited (<25 years), making it difficult to statistically determine return period surge at the downstream boundary of the watershed (Figure 2). In this study, we utilize the longer tidal record located at Galveston Pier 21 in combination with local precipitation gages and the full tropical cyclone data set for the Gulf of Mexico to generate boundary conditions for the Clear Creek Watershed.

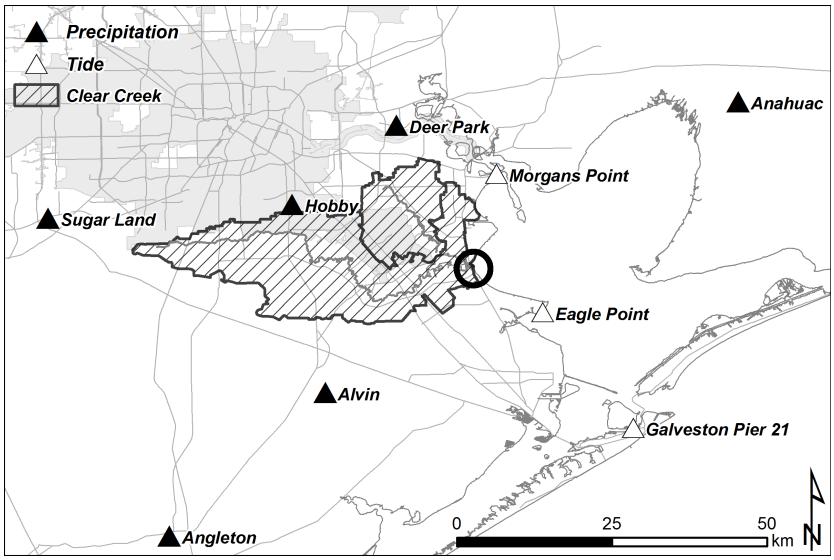

Figure 2: Figure showing the location of the study area on the west side of Galveston Bay and relevant tide and precipitation gages. The circle designates the location where frequency surge information is needed for coastal floodplain studies.

4. Method

The method applied in this analysis consists of several steps shown in the schematic in Figure 3. First, historical tropical cyclone data was collected for the Gulf of Mexico and a database of tropical cyclone characteristics for landfalling storms was created using ArcGIS and Matlab (I). Observed hourly and predicted water levels were collected from a tidal gage and used to calculate maximum daily residual water levels (RWL) near the entrance to Galveston Bay (II). Average daily precipitation was calculated using six rainfall gages surrounding the Clear Creek Watershed (III). Then, the dependence structures between the recorded hurricane characteristics at landfall, peak surge, and cumulative precipitation were analyzed through copulas, and a non-parametric Bayesian network was constructed and validated (IV). The NPBN was used to generate 100,000 synthetic hurricanes which were simulated using an empirical wind set-up model for Galveston Bay to calculate surge at the downstream boundary of the watershed for the synthetic hurricane events (V). Finally, the modeled results were fitted to probability distributions and used to evaluate annual return frequencies for peak storm surge and cumulative precipitation in

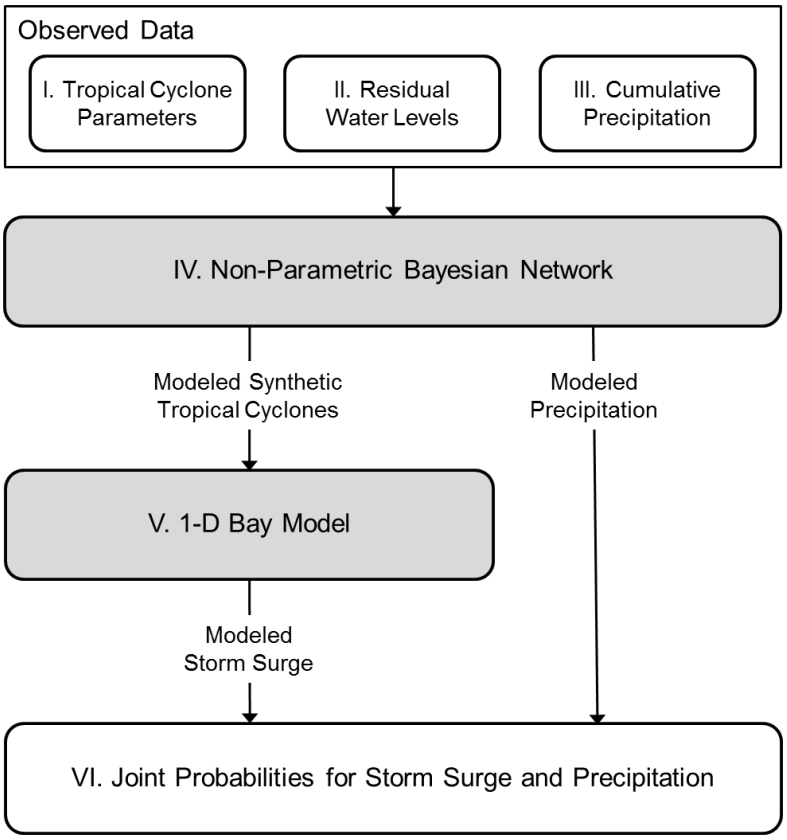

Figure 3: Schematic work flow for generating joint exceedance probabilities of peak storm surge and cumulative precipitation in a coastal watershed.

the watershed (VI). The method is described in more detail in the following sections.

4.1. Data Collection

Tropical Cyclones. Several datasets were combined to generate historical storm characteristics used in this study. The primary dataset is the historical tropical cyclone track information obtained from the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center's International Best Track Archive for Climate Stewardship (IBTrACS) [37]. IBTrACS combines best track data from more than ten meteorological centers worldwide for observed tropical cyclones since 1848. For the purpose of this study, only landfalling tropical cyclones in the Gulf of Mexico have been considered.

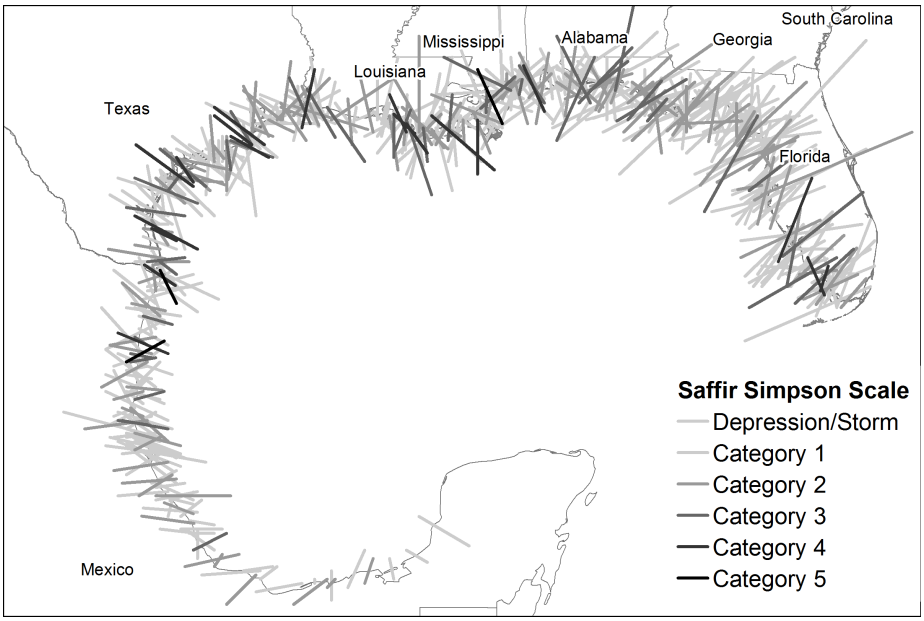

Figure 4: Landfall locations, angle, and intensities of Tropical Cyclones in the Gulf of Mexico (1848-2014).

The IBTrACS dataset was modified in ArcGIS 10.2 to remove all storms that did not enter the Gulf of Mexico or make landfall on the Gulf Coast. For each of the remaining 596 storms, a landfall location was determined where the track intersected the coastline and were assigned cyclone characteristics from one time step - typically six hours - before landfall (Figure 4). The cyclone characteristics obtained from the IBTrACS database include windspeed, pressure, and radius to maximum winds (R_{max}). In addition, angle of approach and forward speed were calculated in Matlab using the latitude-longitude and time-step data for each storm. The landfall location variable was calculated as distance relative to Cancún, Mexico, along the simplified coastline. A summary of available data corresponding to tropical cyclones in the Gulf of Mexico is shown in Table 1.

Storm Surge. Historic water levels were collected at Galveston Pier 21 tide gage (Station ID: 8771450), which is located in the harbor between Galveston and Pelican Islands near the entrance to Galveston Bay (location: $29°18.6'N$, $94°47.5'W$). The Pier 21 gage is maintained by NOAA and is the longest operating tide gage in the Gulf of Mexico. It has a fairly comprehensive record

Table 1: Summary of data collected for tropical cyclones in the Gulf of Mexico (1848-2014).

Variable	Units	Record	No. Obs	Range	Mean	St.Dev.
Windspeed	(kts)	(1851-2014)	596	15-150	59.80	26.73
Pressure	(mb)	(1852-2014)	229	913-1013	981.57	22.95
R_{max}	(km)	(2001-2014)	59	5-150	33.05	25.29
Approach Angle	(deg)	(1851-2014)	596	4.6-359.9	245.45	65.44
Forward Velocity	(mps)	(1851-2014)	596	0.22-23.85	5.62	2.96
Landfall Location	(km)	(1851-2014)	596	568.6-4778.2	3021.2	1046.35
Storm Surge	(m)	(1900-2014)	385	-0.28-4.50	0.20	0.44
Precipitation	(mm)	(1900-2014)	414	0-358.17	28.99	55.01

of hourly water levels beginning January 1, 1908 and extending to December 31, 2013. The long-term linear trend of sea level rise at the gage between 1908 and 2014 was previously studied by NOAA; it was calculated to be 6.34 +/- 0.24 mm/yr based on monthly sea level data and includes corrections for seasonal variations in coastal ocean temperatures, salinities, winds, atmospheric pressures, and ocean currents [38]. The hourly verified water levels collected at the gage were adjusted for sea level rise and are reported herein relative to NAVD88 based on the present (1983-2001) National Tidal Datum Epoch (NTDE) established by the Center for Operational Oceanographic Products and Services (CO-OPS). For the purpose of this study, it is also assumed that the gage location is fairly well-protected from wave activity.

To find the peak storm surge caused by a given tropical cyclone at the gage, the hourly predicted water level (i.e., tide) was subtracted from the SLR-corrected verified water level (i.e., storm tide). For each of the storms that made landfall on the Gulf of Coast, the maximum residual water level (RWL) in the 48 hours surrounding landfall (+/- 48 hours) (when available) was taken as the peak storm surge. This was considered to be effective since no two surge events occur within the same five-day period and this also accounts for the time-lag from any surge generated by storms that make landfall far from the study area. Finally, this time period accounts for the majority of the rainfall associated with

a landfalling event.

In addition, peak water levels for known tropical cyclones were also compared to the maximum water levels reported by NOAA for the top ten events at the gage [39]. For the Unnamed 1915, Unnamed 1919, and Ike 2008 tropical cyclone events, the maximum verified water level did not match the water levels reported by NOAA. For these events, the maximum storm surge was estimated by subtracting the maximum predicted tide level during the period missing from the record from the maximum water level reported by NOAA. This was considered to be a conservative approach. The database was also extended to include the 1900 Galveston Hurricane, a known extreme surge event and the deadliest hurricane in U.S. history [40]. Since no predicted tides are available for this date, the maximum water level is used to represent storm surge. Based on the hurricane record, no significant events occurred in the Houston-Galveston region between 1900 and 1908 making this a reasonable extension of the data set.

Cumulative Precipitation. Daily precipitation was collected from the National Climatic Data Center (NCDC) for six gages in the vicinity of the Clear Creek watershed (Figure 2). The daily precipitation over the Clear Creek Watershed between January 1, 1900 and December 31, 2013 was calculated using the Thiessen polygon method to spatially average the rainfall over the watershed for the combination of available gages on each day in the record. To identify cumulative precipitation corresponding to peak surge, the five-day cumulative precipitation surrounding the date of hurricane landfall was calculated. The decision to use the five-day cumulative rainfall was made based on the average time to peak of the hyetograph for rainfall events during hurricanes. For the majority of events, greater than 90% of the total cumulative rainfall occurs during the first 72 hours after landfall when analyzed over a 10-day period. No lag was applied to the rainfall data as the response time of the watershed is nearly instantaneous (i.e., peak daily rainfall coincides with peak daily runoff within one day) and, for the purpose of this study, only the cumulative rainfall near the time of peak surge is of interest.

4.2. NPBN Construction & Validation

In this section, we provide further details about the construction and validation of the NPBN applied in this study. The bivariate copulas giving rise to the dependence structure under investigation are assumed to come from the Gaussian (Normal) copula and have the following distribution function:

$$C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), (u,v) \in [0,1]^2 \tag{6}$$

where Φ^{-1} is the inverse Gaussian cumulative distribution and ρ is the (conditional) product moment correlation. The advantage of using the Gaussian Copula to describe the joint distributions in the model is the rapid calculation of the synthetic storm data set in subsequent steps.

The Gaussian copula assumption was tested against two alternative copulas (i.e., Gumbel and Clayton) using the semi-correlation ([24]) and Cramér-von-Mises tests described in [41] and provided for reference in Appendix A.2. Together, these three copula models were used to investigate the ranges of dependence that are usually observed in environmental data, (See discussion of tail dependencies in Appendix A.1). The results of the goodness-of-fit tests for variable pairs where the absolute value of the correlation coefficient, ρ, exceeds 0.1 are shown in Table 2. In general, the tests show that the Gaussian copula is an acceptable approximation indicated by low Cramér-von-Mises statistic values and small differences in semi-correlations (bolded values in Table 2). For those variable pairs in which the Gaussian copula was not the best fit, the difference in the Cramér-von-Mises statistic with respect to the Gaussian copula is sufficiently small. Additional discussion of the results of these tests is included in Appendix A.2.

After validating the application of the Gaussian Copula to describe the dependence relationships between the variables, the NPBN was constructed in Matlab and the UniNet software package was used to visualize the model [42]. To maximize the number of samples used to build the network (i.e., 302), the NPBN was built using the empirical distributions and joint data for six variables: wind, forward velocity, angle, landfall location, storm surge, and precipitation. Arcs were drawn between nodes where absolute value of the (conditional) correlation coefficient, ρ, exceeded 0.1.

Table 2: Goodness-of-Fit tests. Semi-correlation and Cramér-von-Mises statistic (CM_n) for pairs where $|\rho| > 0.1$. ρ_{NW}, ρ_{NE}, ρ_{SE}, and ρ_{SW} are the correlation coefficients calculated in the four quadrants (e.g., northwest, northeast, southeast, and southwest), respectively, for each pair of variables. Relevant semi-correlations and lowest CM_n values are bolded.

		ρ	ρ_{NW}	ρ_{NE}	ρ_{SE}	ρ_{SW}	CM_n **(Ga)**	CM_n **(Gu)**	CM_n **(Cl)**
Windspeed	Angle	**-0.11**	**0.16**	-0.35	**0.14**	-0.06	**0.31**	0.44	0.43
Windspeed	Surge	**0.36**	0.28	**0.31**	0.12	**-0.07**	0.41	**0.40**	0.91
Velocity	Angle	**0.39**	-0.21	**0.38**	0.10	**0.29**	0.21	**0.18**	0.46
Velocity	Landfall Loc.	**0.26**	-0.12	**0.44**	0.06	**0.00**	0.33	**0.25**	0.66
Angle	Landfall Loc.	**0.71**	-0.26	**0.32**	0.35	**0.55**	**0.24**	0.44	0.86
Landfall Loc.	Precipitation	**-0.13**	**0.34**	-0.18	**-0.28**	-0.04	1.16	**1.00**	1.76
Surge	Precipitation	**0.38**	0.33	**0.36**	-0.12	**-0.02**	0.45	**0.28**	1.16

In addition, radius to maximum winds was added to the NPBN as a user-defined random variable (UDRV). Radius to maximum winds (R_{max}) has only become part of the tropical cyclone record since the early 2000s when its importance was recognized for predicting hurricane intensity and damage. Therefore, there are only 59 records of R_{max} in the IBTrACS database for the Gulf of Mexico. We used these records to establish relationships between R_{max} and the other variables and found correlations ($|\rho| > 0.1$) between R_{max} and maximum wind speed and landfall location. R_{max} was added to the NPBN model as a continuous UDRV with a log-normal parametric distribution. The NPBN structure was validated using the tests described in the supplementary material and is shown in Figure 5.

Because the empirical wind set-up model for the bay requires that all storms make landfall perpendicular to the coast, the NPBN is conditioned on a fixed angle of 235 degrees relative to due east. After conditioning, the remaining probability distributions are updated and the NPBN is sampled to generate 100,000 synthetic storms that were then input into the empirical wind set-up model. Two examples demonstrating the flexibility of the NPBN are included in Appendix A.4.

4.3. Probabilistic Empirical Wind Set-up Model

Of particular interest in this study is the return period surge and precipitation for the Clear Creek watershed. To obtain surge heights at the downstream boundary of the Clear Creek watershed, we use a simplified 1-D, empirical wind

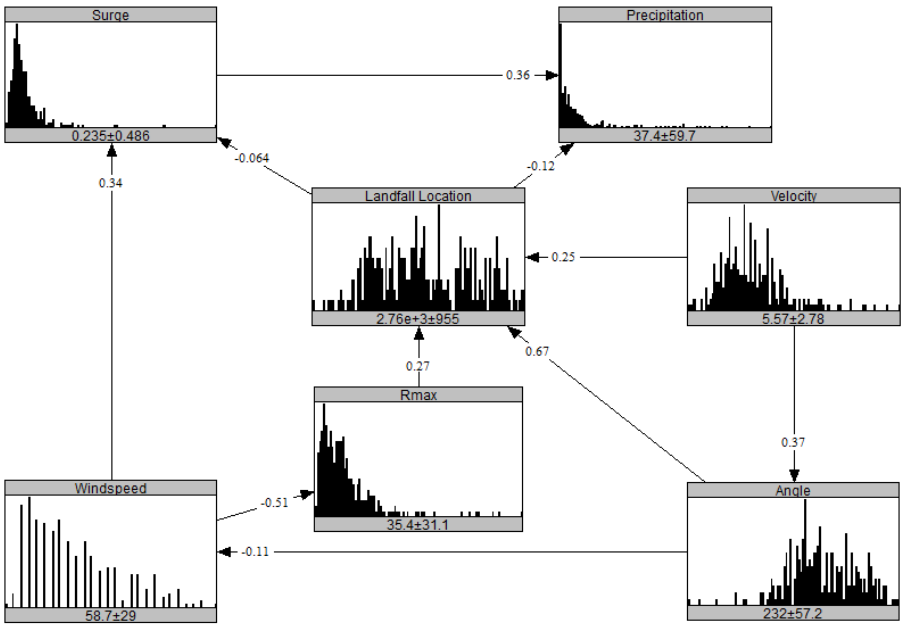

Figure 5: The Bayesian network (BN) structure for tropical cyclones impacting the Galveston Bay Region. The nodes are presented as histograms and the mean and standard deviations for each variable are given. The (conditional) rank correlation coefficients are shown on each arc.

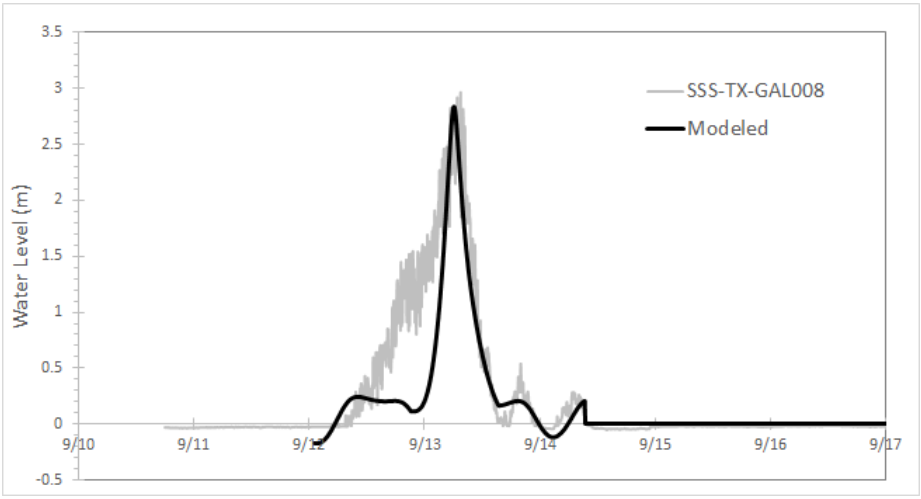

Figure 6: Comparison of modeled output with observed data at USGS Gage SSS-TX-GAL008 for Hurricane Ike (2008).

set-up model for the bay as proposed by [43]. In this model, a parametric hurricane wind field is built using values of pressure and R_{max}. Forward velocity and landfall location are used to translate the synthetic hurricane along a track perpendicular to the coast. Storm surge at the coast is calculated by solving the one-dimensional depth integrated shallow water equations, and storm surge within the bay is calculated based on a parametric relation between wind set-up and storm surge at the open-coast. Hindcasts of historic storms show that the model provides a reasonable estimate of storm surge in the bay within 0.5 meters [43]. A comparison between observed and model results for Hurricane Ike (2008) is plotted in Figure 6.

To estimate the annual return periods for storm surge at the downstream boundary of Clear Creek, we fit the output from the probabilistic model to the three-parameter generalized Pareto distribution (GPD):

$$y = f(x|k,\sigma,\theta) = (\frac{1}{\sigma})(1 + k\frac{(x-\theta)}{\sigma})^{-1-\frac{1}{k}} \tag{7}$$

where k is the shape parameter, σ is the scale parameter, and θ is the location parameter. This distribution is often used to model the tails of extreme data and has been applied in previous studies of storm surge [10]. The distribution has three forms: the tail increases exponentially ($k = 0$); the tail decreases as a polynomial ($k > 0$); the tail has a finite limit ($k < 0$). For the entire data set,

the best fit was a GPD distribution where $k = 0$. However, to avoid completely unrealistic estimates of surge within Galveston Bay, we chose to fit the upper tail of the data where surge exceeded a minimum threshold of 0.55 m, which resulted in a negative shape parameter and finite limit for storm surge at our location of interest. The probable maximum surge was then evaluated using the equation:

$$x = \theta - \frac{\sigma}{k} \tag{8}$$

The GPD distribution was adjusted for the annual return frequency of tropical cyclones exceeding the minimum threshold in Galveston Bay using the Poisson estimator, $\hat{\lambda}$, based on the average annual number of tropical cyclones.

5. Simulation Results

The results from the NPBN-based approach were visually compared to annual return frequency graphs from two recently published studies that used the ADvanced CIRCulation (ADCIRC) Model to analyze storm surge in the Galveston Bay region (Figure 7). Both studies apply the Joint Probability Method-Optimum Sampling (JPM-OS) approach to reduce the number of representative synthetic storms needed for simulation [44, 13]. However, even with the application of the JPM-OS approach, probabilistic modeling using high-resolution storm surge models is computationally demanding, requiring large computer clusters to run [10, 12, 13].

As seen in Figure 7, the NPBN-approach produces lower predictions of surge in the low-frequency region as compared to other studies; however, there is general agreement on the 50-year return period surge. It is important to note that neither the FEMA [13] study, nor the study by Ebersole et al. [44] discuss the validity or accuracy of their results with respect to historical data. When compared to the available observed data at Eagle Point (1994-2014), our model shows similar limitations to existing studies: it over-predicts storm surge in the high-frequency region of the graph and potentially under-predicts in the low-frequency portion of the graph. However, the frequency of observed events is not well-understood due to lack of historical data in Galveston Bay. Most importantly, we see here that there is little consensus on the return frequency values for surge in the Galveston Bay region even when using high-resolution,

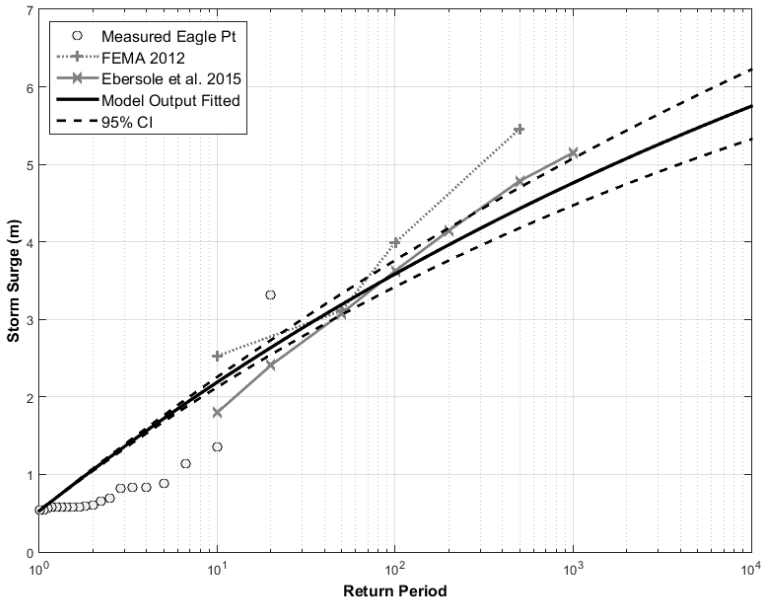

Figure 7: Comparison of modeled output at the downstream boundary of Clear Lake with existing studies published by Ebersole et al. [44] and FEMA [13]. The empirical distribution of available storm surge data at Eagle Point (1994-2014) is shown in red for comparison.

computationally expensive models. While the two existing studies generally agree on a value for the 1/50 year storm surge, there is no consensus for other return periods, making it difficult to establish a benchmark against which to verify our model results.

Figure 8 shows the joint probability contours for peak surge and cumulative precipitation. To calculate the joint probability of peak surge and precipitation, the variables were fit to their marginal distributions and the iso-probability contours were mapped in the probability density space. The contours for the 10-, 50-, 500-, and 1000-year recurrence intervals are shown as black dashed lines and the contour for the 100-year event is shown as a black solid line. For hazard mitigation in the U.S., we are particularly interested in the 1% annual chance of occurrence for storm surge and precipitation. For comparison, the contours are plotted against the available observed data at Eagle Point (1999-2014), the closest tide gage to Clear Creek. Based on the results, we see that when considering both storm surge and precipitation, Hurricane Ike exceeded a 100-year event.

6. Discussion

In this section, we discuss the limitations of the methodology and provide suggestions for future work. We acknowledge that there are multiple sources of uncertainty in this study and divide them into three categories: (1) uncertainty due to characterization of historical data and approximation of statistical distributions; (2) accumulated errors and assumptions in the model; and (3) uncertainty due to potential changes in climate and topology over time that are not accounted for in this approach.

First, while the BN configuration presented in this study was a reasonable fit given the data available, it may not be the only solution possible. Previous studies have noted the importance of tropical cyclone characteristics such as radius to tropical storm winds, R_{TS}, and tropical cyclone intensity prior to landfall in predicting surge height [12, 10]. For example, Resio et al. [10] argue that the use of landfall characteristics tends to underestimate tropical cyclone intensities offshore since storms typically loose energy as they make landfall. However, in many cases, historical data pertaining to offshore characteristics is

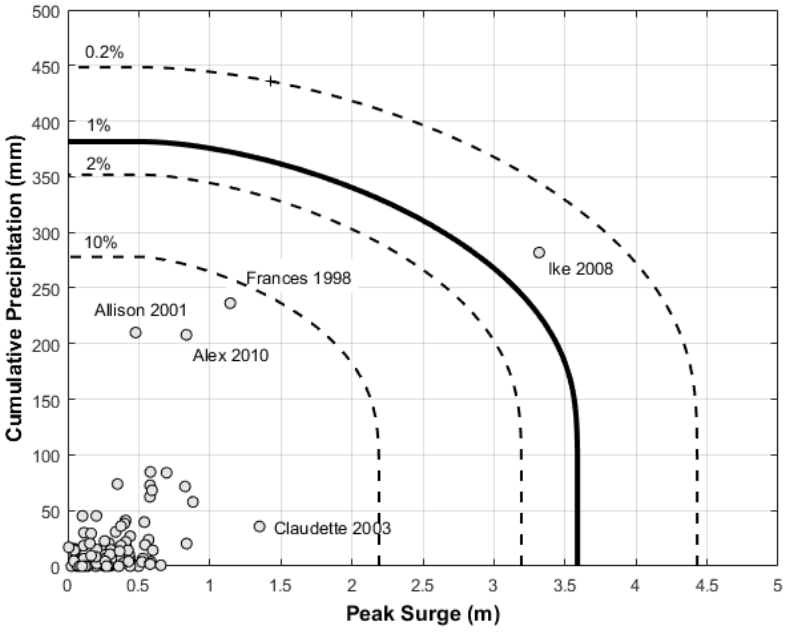

Figure 8: Joint probability plot for cumulative precipitation (mm) and peak storm surge (m). The 10-, 50-, 500-, and 1000-year contours are shown as black dashed lines. The 100-year contour is shown as a solid black line. Observed values at Eagle Point (surge) and in Clear Creek (precipitation) are shown in red.

difficult to obtain or non-existent. Future work should explore the uncertainty associated with using landfall characteristics to build the Bayesian network and strive to incorporate additional data from offshore observations when available.

Second, for simplicity, we assumed that all of the dependence structures in the Bayesian network can be described using the Gaussian copula. However, based on the statistical tests described in Section 4.2, we observed that there are some bivariate distributions that display non-Gaussian behavior. Most significantly, the results indicate that surge and precipitation may exhibit upper tail dependence that may be better described using the Gumbel copula. Using the Gaussian copula to describe this dependence structure may have led to conservative estimates of combinations of peak surge and precipitation for extreme (i.e., low-frequency) events. Future work should further explore the relationships between these two variables.

In addition, we apply a simplified empirical wind-setup model to estimate surge within the bay system. This model has noted weaknesses, including: oversimplification of the coastline and bay bathymetry, assumption that all storms make landfall perpendicular to the coast, and constant wind fields prior to landfall. Despite these limitations, the lack of consensus between the different studies using the JPM-OS approach for ADCIRC and the comparison of our results with historical data lead us to believe that our method is a reasonable alternative to computationally-expensive modeling. While we recommend that the simplified model be further improved, we find that proposed approach is appropriate for estimating return period surge, especially in areas with limited historical data when considering computational time and processing power.

Third, while we corrected for sea level rise and tested the surge and precipitation for trends in the data, changes in topography and bathymetry over the previous century (1900-2014) may have impacted the integrity of the water levels recorded at Galveston Pier 21. Human and nature-induced changes to the barrier islands and Galveston Bay, such as the construction of a concrete seawall on Galveston Island, dredging of the Houston Ship Channel, and long-term and episodic erosion of the barrier islands, may have changed the response of the system to tropical cyclones over time.

In future work, we also recommend investigating time-dependent rainfall and

surge intensities. For example, during Hurricane Carla (1961) surge reached 2.42 m and remained above 0.67 m for multiple days. In this case, a longer period of rainfall might be of consequence because the surge would not allow the rainfall-runoff to exit the watershed, thus causing an exacerbated backwater profile. In contrast, during Hurricane Claudette (1979) the surge was lower; nevertheless, the rainfall spanned multiple days causing extreme flooding in the watershed. Future studies should incorporate the timing and intensity of rainfall with the timing and intensity of surge to account for worst-case scenario flood events for the watershed.

7. Conclusion

In this paper, we present a method for obtaining an initial approximation of joint exceedance probabilities for peak surge and cumulative precipitation in the Clear Creek Watershed in Southeast Texas. We present a statistical model based on historical hurricanes in the Gulf of Mexico. The model was built using the IBTrACS storm database and gage records for extreme water levels and precipitation in the study region, and was validated for historical events where storm surge exceeded 0.5 meters at the coast. The results indicate that the modeling scheme provides similar results for return frequency approximations when compared to high-resolution coastal models previously applied to the Galveston Bay system. In addition, the model provides a first estimate of probable precipitation associated with return-period surge values and, to our knowledge, this is the first study of joint probabilities for precipitation and surge from hurricanes in the Gulf Coast region. The advantage of this methodology is that it is flexible and computationally efficient. Nevertheless, large uncertainties remain due to assumptions made in the dependence structure of the NPBN (i.e., Gaussian copula), simplifications in the empirical wind set-up model (i.e., average depth and circular bay), and quality of observed data.

While the results from this study provide valuable information for the Clear Creek Watershed in the Houston-Galveston region, future work will focus on expanding the model to include observed data for more watersheds in region and for other sections of the Gulf Coast. Such results will provide critical information for flood risk assessment in vulnerable communities along the U.S. Gulf

Coast. The model can also be applied to event-response planning since it can be used as a probabilistic prediction tool and has potential real-time applications (see Appendix A.4). As the frequency and intensity of compound flood events increase under changing climate, and urban settlements grow around coastal cities, it will become increasingly important to fully and accurately assess the risk of compound flooding in coastal areas.

8. Acknowledgments

This work was supported by the Netherlands America Foundation/Fulbright Fellowship for Water Management and the NSF PIRE Grant No. OISE-1545837. The authors would also like to thank K. Stoeten for sharing the probabilistic model and S.N. Jonkman for his comments regarding this research. We would also like to acknowledge two anonymous reviewers and thank them for their valuable comments which helped to improve this manuscript.

References

[1] UN-Oceans, UN Atlas of the Oceans: Human Settlements on the Coast (2016).
URL http://www.oceansatlas.org/servlet/CDSServlet?status=ND0xODc3JjY9ZW4mMzM9KiYzNz1rb3M{~}

[2] W. E. Highfield, S. A. Norman, S. D. Brody, Examining the 100-Year Floodplain as a Metric of Risk, Loss, and Household Adjustment, Risk Analysis 33 (2) (2013) 186–191. doi:10.1111/j.1539-6924.2012.01840.x.

[3] Gulf of Mexico at a Glance: A Second Glance, Tech. rep., National Oceanic and Atmospheric Administration (NOAA) (2011).
URL
http://gulfofmexicoalliance.org/pdfs/gulf{_}glance{_}1008.pdf

[4] T. A. Birkland, R. J. Burby, D. Conrad, H. Cortner, W. K. Michener, River Ecology and Flood Hazard Mitigation, Natural Hazards Review 4 (1) (2003) 46–54. doi:10.1061/(ASCE)1527-6988(2003)4:1(46).

[5] FEMA, Coastal Flood Risk Study Process (2015).
URL https://www.fema.gov/coastal-flood-risk-study-process

[6] H. Apel, A. H. Thieken, B. Merz, G. Blöschl, Flood risk assessment and associated uncertainty, Natural Hazards and Earth System Science 4 (2) (2004) 295–308. doi:10.5194/nhess-4-295-2004.

[7] R. E. Morss, O. V. Wilhelmi, M. W. Downtown, E. Gruntfest, Flood risk, uncertainty, and scientific information for decision making: lessons from an interdisciplinary project, Bulletin of the American Meteorological Society 86 (11) (2005) 1593–1601. doi:10.1175/BAMS-86-11-1593.

[8] R. Hirsch, T. Cohn, W. Kirby, What does the 1% flood standard mean? Revisiting the 100-year flood, in: Gilbert F. White National FLood Policy Forum: Reducing Flood Losses : Is the 1 % Chance (100-year) Flood Standard Sufficient ?, National Academies Disasters Roundtable, Washington, D.C., 2004, pp. 117–122.

[9] Meteorological criteria for standard project hurricane and probable maximum hurricane windfields, gulf and east coasts of the United States NOAA Technical Report NWS 23, Tech. rep., National Oceanic and Atmospheric Administration (NOAA), Washington, D.C. (1979).

[10] D. T. Resio, J. Irish, M. Cialone, A surge response function approach to coastal hazard assessment part 1: basic concepts, Natural Hazards 51 (1) (2009) 163–182. doi:10.1007/s11069-009-9379-y.

[11] F. Ho, J. Su, K. Hanevich, R. Smith, F. Richards, Hurricane Climatology for the Atlantic and Gulf Coasts of the United States, Tech. rep., National Oceanic and Atmospheric Administration (NOAA), Silver Spring, MD (1987).

[12] J. L. Irish, D. T. Resio, M. a. Cialone, A surge response function approach to coastal hazard assessment. Part 2: Quantification of spatial attributes of response functions, Natural Hazards 51 (1) (2009) 183–205. doi:10.1007/s11069-009-9381-4.

[13] Flood Insurance Study: Galveston County, Texas, and incorporated areas, Tech. rep., Federal Emergency Management Agency (FEMA) (2012).

[14] A. Sebastian, J. Proft, J. C. Dietrich, W. Du, P. B. Bedient, C. N. Dawson, Characterizing hurricane storm surge behavior in Galveston Bay using the SWAN+ADCIRC model, Coastal Engineering 88 (2014) 171–181. doi:10.1016/j.coastaleng.2014.03.002.
URL http://linkinghub.elsevier.com/retrieve/pii/S0378383914000556

[15] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers, San Mateo, 1988.

[16] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks, 1st Edition, Springer Publishing Company, Incorporated, 2007.

[17] D. Kurowicka, R. Cooke, Uncertainty analysis with high dimensional dependence modelling, John Wiley & Sons Ltd., 2006.

[18] A. Hanea, O. Morales Napoles, D. Ababei, Non-parametric Bayesian networks: Improving theory and reviewing applications, Reliability Engineering & System Safety 144 (2015) 265–284. doi:10.1016/j.ress.2015.07.027.

[19] R. Hanea, A., Kurowicka, D., Cooke, Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets, Quality and Reliability Engineering International 22 (2006) 709–729. doi:10.1002/qre.808.

[20] D. Paprotny, O. Morales-Napoles, A Bayesian Network for extreme river discharges in Europe, in: Podofillini (Ed.), Safety and Reliability of Complex Engineered Systems, Taylor & Francis Group, London, 2015, pp. 4303–4311.

[21] M. Sklar, Fonctions de répartition à n dimensions et leurs marges, Ph.D. thesis, Université Paris 8 (1959).

[22] C. Genest, A.-C. Favre, Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask, Journal of Hydrologic Engineering 12 (4) (2007) 347–368. doi:10.1061/(ASCE)1084-0699(2007)12:4(347).

[23] R. B. Nelson, An Introduction to Copulas, 2nd Edition, Springer, 2007.

[24] H. Joe, Dependence Modeling with Copulas, Chapman & Hall/CRC, London, 2014.

[25] STAHY, References on Copula Function topic (2016). URL http://www.stahy.org/Activities/STAHYReferences/ReferencesonCopulaFunctiontopic/tabid/78/Default.aspx

[26] B. Fernandez, D. Salas, Return Period and Risk of Hydrologic Events, I: Mathematical Formulation, Journal of Hydrologic Engineering October (1) (1999) 297–307.

[27] L. Zhang, V. P. Singh, Bivariate rainfall frequency distributions using Archimedean copulas, Journal of Hydrology 332 (1-2) (2007) 93–109. doi:10.1016/j.jhydrol.2006.06.033.

[28] A. Bárdossy, G. Pegram, Copula based multisite model for daily precipitation simulation, Hydrology and Earth System Sciences Discussions 6 (3) (2009) 4485–4534. doi:10.5194/hessd-6-4485-2009.

[29] A. Arns, T. Wahl, I. D. Haigh, J. Jensen, Determining return water levels at ungauged coastal sites: a case study for northern Germany, Ocean Dynamics 65 (4) (2015) 539–554. doi:10.1007/s10236-015-0814-1.

[30] P. H. van Gelder, C. V. Mai, W. Wang, G. Shams, M. Rajabalinejad, M. Burgmeijer, Data management of extreme marine and coastal hydro-meteorological events, Journal of Hydraulic Research 46 (2008) 191–210. doi:10.1080/00221686.2008.9521954.

[31] T. Wahl, J. Jensen, C. Mudersbach, a Multivariate Statistical Model for Advanced Storm Surge Analyses in the North Sea, Coastal Engineering Proceedings 1 (2011) 1–12. doi:10.9753/icce.v32.currents.19.

[32] T. Wahl, C. Mudersbach, J. Jensen, Assessing the hydrodynamic boundary conditions for risk analyses in coastal areas: a multivariate statistical approach based on Copula functions, Natural Hazards and Earth System Science 12 (2) (2012) 495–510. doi:10.5194/nhess-12-495-2012.

[33] J. C. Trepanier, H. F. Needham, J. B. Elsner, T. H. Jagger, Combining Surge and Wind Risk from Hurricanes Using a Copula Model: An Example from Galveston, Texas, The Professional Geographer 67 (1) (2015) 52–61. doi:10.1080/00330124.2013.866437.
URL http://www.tandfonline.com/doi/abs/10.1080/00330124.2013.866437

[34] HGAC, Personal Communication (2011).

[35] NHC, Tropical Cyclone Climatology (2015).
URL http://www.nhc.noaa.gov/climo/

[36] T. Wahl, S. Jain, J. Bender, S. D. Meyers, M. E. Luther, Increasing risk of compound flooding from storm surge and rainfall for major US cities, Nature Climate Change (July) (2015) 1–6. doi:10.1038/NCLIMATE2736.

[37] K. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, C. J. Neumann, The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data, Bulletin of the American Meteor. Society 91 (2010) 363–376. doi:10.1175/2009BAMS2755.1.

[38] NOAA, Sea Level Trends for 8771450, Galveston Pier 21 TX (2013). URL https://tidesandcurrents.noaa.gov/sltrends/sltrends{_}station.shtml?stnid=8771450

[39] Center for Operational Oceanographic Products and Services (CO-OPS), Top Ten Highest Water Levels for long-term stations in meters above MHHW, Tech. rep., National Oceanic and Atmospheric Administration (NOAA), Washington, D.C. (2014).

[40] E. S. Blake, E. J. Gibney, The Deadliest, Costliest, and Most Intense United States Tropical Cyclones from 1851-2010 (and other frequently requested hurricane facts) NOAA Technical Memorandum NWS NHC-6, Tech. rep., National Hurricane Center (NHC), Miami, FL (2011). doi:10.1073/pnas.0703993104.

[41] C. Genest, R. Bruno, D. Beaudoin, Goodness-of-fit tests for copulas : A review and a power study, Insurance: Mathematics and Economics 44 (2009) 199–213. doi:10.1016/j.insmatheco.2007.10.005.

[42] O. Morales-Nápoles, D. Worm, P. van den Haak, A. Hanea, W. Courage, S. Miraglia, Reader for course: Introduction to Bayesian Networks, TNO, Delft, the Netherlands, 2013.

[43] K. Stoeten, Hurricane Surge Risk Reduction For Galveston Bay, Msc thesis, Delft Technical University (2013).

[44] B. A. Ebersole, T. C. Massey, D. L. Hendon, T. W. Richardson, R. W. Whalin, Interim Report - Ike Dike Concept for Reducing Hurricane Storm Surge in the Houston- Galveston Region Jackson State University, Tech. rep., Jackson State University, Jackson (2015).

[45] A.-C. Favre, S. Adlouni, L. Perreault, N. Thiemonge, B. Bobee, Multivariate hydrological frequency analysis using copulas, Water Resources Research 40 (2004) 1–12. doi:10.1029/2003WR002456.

[46] H. Joe, Multivariate Models and Multivariate Dependence Concepts, Chapman & Hall/CRC Press, 1997.

[47] D. Berg, K. Aas, Models for construction of multivariate dependence a comparison study, The European Journal of Finance 15 (7-8) (2009) 639–659. doi:10.1080/13518470802588767.

Appendix A. Supplementary Material

This appendix provides additional information pertaining to the statistical tests referenced in this paper. First we present an analysis of the historical data collected for the paper. Then, we describe the copulas that were tested for the non-parametric Bayesian network (NPBN), the goodness-of-fit tests, and a justification of the choice of the Gaussian copula. Finally, we provide the equations used to validate the structure of the NPBN and present the results.

Appendix A.1. Copulas

There are many families of copula models available (see [24]), however, Gaussian and Archimedean Copulas, such as Clayton, Frank, and Gumbel, are most often used for hydrologic analysis [22, 45, 32]. In our study we compare the performance of three of the most popular copulas (Gaussian, Gumbel, Clayton) in order to determine whether certain characteristics modeled by these families are present in the bivariate distributions for the variables included in the NPBN model for Galveston Bay.

Gaussian Copula. Equation A.1 presents the Gaussian copula

$$C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), (u,v) \in [0,1]^2 \tag{A.1}$$

where ρ is the (conditional) product moment correlation, Φ is the bivariate Gaussian cumulative distribution, and Φ^{-1} is the inverse of the 1-D Gaussian cumulative distribution function (cdf).

Gumbel Copula. Equation A.2 presents the Gumbel copula

$$C_\delta(u,v) = \exp[-([-log(u)]^\delta + [-log(v)]^\delta)^{(1/\delta)}], \delta \geq 1 \tag{A.2}$$

where the copula is parameterized by δ.

Clayton Copula. Equation A.3 presents the Clayton copula

$$C_\alpha(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\alpha}, \alpha \in [-1, \infty] \tag{A.3}$$

where the copula is parameterized by α.

Together, these three copulas model cover a range of tail dependencies usually observed in the data, where the upper tail dependence coefficient λ_U for random variables X_i and X_j is defined as

$$\lambda_U = \lim_{u \to 1} P(X_i > F_{X_i}^{-1}(u) | X_j > F_{X_j}^{-1}(u)) = \lim_{u \to 1} P(U > u | V > u) \tag{A.4}$$

Lower tail dependence is defined similarly, but for the lower quadrant of the joint distribution. In the case of the three copulas presented herein, the Gaussian copula exhibits no tail dependence (i.e., $\lambda_U = 0$), the Gumbel copula exhibits upper tail dependence (i.e., $\lambda_U = 1 - 2^{1/\delta}$), and the Clayton copula exhibits lower tail dependence (i.e., $\lambda_U = 2^{1/\alpha}$).

A complete introduction to copulas and multivariate models can be found in the books written by [46] and [23], and a review of applications and types of copula models can be found in [47].

Appendix A.2. Goodness-of-Fit Tests for Copulas

In this study we apply two tests to determine whether certain characteristics of the copulas under investigation are present in the data: semi-correlations and the Crámer-von-Mises statistic.

Semi-correlations. Semi-correlations are used to test how well a chosen copula fits the tail(s) of a data set [24]. We compare the overall correlation, ρ, for the normal transform of the original data, Z_λ, to the Pearson's product moment correlation coefficient (i.e. semi-correlation) in the "tail" quadrants. That is upper right (ρ_{ne}) and lower left (ρ_{sw}) for positively correlated data, and upper left (ρ_{nw}) and lower right (ρ_{se}) for negatively correlated data. In general, if the absolute value of the semi-correlation is greater than the overall correlation, then the data is considered to be tail dependent.

For positively correlated variables X_i and X_j, the semi-correlations are computed as follows:

$$\rho_{ne} = \rho(Z_i, Z_j | Z_i > 0, Z_j < 0) \tag{A.5}$$

$$\rho_{sw} = \rho((Z_i, Z_j | Z_i > 0, Z_j < 0) \tag{A.6}$$

where (Z_i, Z_j) are the standard normal transforms of X_i and X_j. Similarly, if the variables X_i and X_j are negatively correlated, the semi-correlations are:

$$\rho_{nw} = \rho(Z_i, Z_j | Z_i > 0, Z_j > 0) \tag{A.7}$$

$$\rho_{se} = \rho((Z_i, Z_j | Z_i < 0, Z_j < 0) \tag{A.8}$$

Cramér-von Mises statistic. The Cramér-von Mises (CM) Statistic performs as a "blanket test" for the entire data set [41]. The CM statistic is the sum of the squared differences between the empirical copula and the parametric estimate (e.g., Gaussian, Gumbel, or Clayton) for a given number of samples. In general, the copula with the lowest CM Statistic is the best estimate for the data set.

The test statistic for a sample of length n is computed as

$$CM_n(\mathbf{u}) = \sum\nolimits_{|\mathbf{u}|} \{C_{\hat{\theta}_n}(\mathbf{u}) - B(\mathbf{u})\}^2, \mathbf{u} \in [0,1]^2 \tag{A.9}$$

where $B(\mathbf{u}) = \sum 1(U_i \leq \mathbf{u})$ is the empirical copula and $C_{\hat{\theta}_n}(\mathbf{u})$ is a parametric copula with parameter $\hat{\theta}_n$ estimated from the sample.

The results from both tests applied to the variables in our study are presented in Table A.3 and plots for interesting cases are shown in Figures A.9(a-d). The results indicate that the Gaussian copula behaves well for many of the variable pairs (i.e., 7/15) indicated by low CM_n statistic values and small differences in the semi-correlations (see bolded values in TableA.3). The CM_n statistic indicates that the Gumbel copula is the best fit for six pairs and that the Clayton copula is the best fit for two pairs. In the case of five of the eight pairs in which the Gaussian copula is not the best fit, the difference in the CM_n statistic is small with respect to the Gaussian copula (≤ 0.03) (i.e., Windspeed-Velocity, Windspeed-Landfall Location, Windspeed-Surge, Velocity-Angle).

However, for four variable pairs (i.e., Velocity-Landfall Location, Distance-Surge, Distance-Precipitation, and Surge-Precipitation), the differences between the CM_n Statistic for the Gaussian and Gumbel copulas is larger (0.08-0.17). In general, these variable pairs display only slight upper tail dependence and the Gaussian copula is still a valid assumption. The four variable pairs are plotted in Figure A.9. Interestingly, surge and precipitation display upper tail dependence characteristic of the Gumbel copula and in future studies this should be further evaluated.

Table A.3: Semi-correlations and Cramér-von Mises statistics (CM_n) for all variable pairs analyzed in the Bayesian network. Relevant semi-correlations and lowest CM_n values are bolded.

		ρ	ρ_{NW}	ρ_{NE}	ρ_{SE}	ρ_{SW}	CM_n(**Ga**)	CM_n(**Gu**)	CM_n(**Cl**)
Windspeed	Velocity	**0.04**	0.15	**-0.03**	0.14	**-0.02**	0.24	0.31	**0.23**
Windspeed	Angle	**-0.11**	**0.16**	-0.35	**0.14**	-0.06	**0.31**	0.44	0.43
Windspeed	Landfall Loc.	**-0.07**	**-0.08**	-0.10	**0.25**	-0.09	0.28	0.44	**0.25**
Windspeed	Surge	**0.36**	0.28	**0.31**	0.12	**-0.07**	0.41	**0.40**	0.91
Windspeed	Precipitation	**0.00**	0.25	**0.12**	-0.09	**-0.03**	**0.30**	0.36	0.39
Velocity	Angle	**0.39**	-0.21	**0.38**	0.10	**0.29**	0.21	**0.18**	0.46
Velocity	Landfall Loc.	**0.26**	-0.12	**0.44**	0.06	**0.00**	0.33	**0.25**	0.66
Velocity	Surge	**0.11**	-0.14	**-0.12**	0.00	**-0.11**	**0.16**	0.26	0.17
Velocity	Precipitation	**-0.07**	**-0.13**	-0.15	**-0.34**	0.04	**0.15**	0.18	0.15
Angle	Landfall Loc.	**0.71**	-0.26	**0.32**	0.35	**0.55**	**0.24**	0.44	0.86
Angle	Surge	**0.01**	0.19	**-0.30**	0.16	**-0.11**	**0.34**	0.35	0.34
Angle	Precipitation	**-0.03**	**0.10**	-0.19	**-0.24**	-0.01	**0.35**	0.35	0.37
Landfall Loc.	Surge	**-0.09**	**0.54**	-0.18	**-0.09**	0.11	1.99	**1.82**	2.26
Landfall Loc.	Precipitation	**-0.13**	**0.34**	-0.18	**-0.28**	-0.04	1.16	**1.00**	1.76
Surge	Precipitation	**0.38**	0.33	**0.36**	-0.12	**-0.02**	0.45	**0.28**	1.16

Appendix A.3. Validation of the Network

In this study, we utilize two tests to validate the network construction: the validation tests presented by [18] and the d-calibration score presented by [42]. The objective of these tests is to determine whether the normal copula hypothesis is valid for the network and whether the network structure is appropriate.

To do so, it is necessary to calculate the "closeness" between the determinant of the empirical rank correlation matrix (DER) and the determinant of the empirical normal rank correlation matrix (DNR), and the "closeness" between the DNR and the determinant for a non-parametric Bayesian network based on normal copulas (DBN). The DER is calculated by transforming the marginals to uniforms (normalizing) and then calculating the product moment correlation of the transformed variables. Similarly, the DNR is obtained by transforming the marginals to standard normal and then transforming the product moment correlations to rank correlations using the formula:

$$r(X_i, X_j) = \frac{6}{\pi} arcsin\left(\frac{\rho(X_i, X_j)}{2}\right) \tag{A.10}$$

If the DER is within the 90% confidence bounds of the DNR (and, similarly, the DNR is within the 90% of the DBN), the network structure is considered to

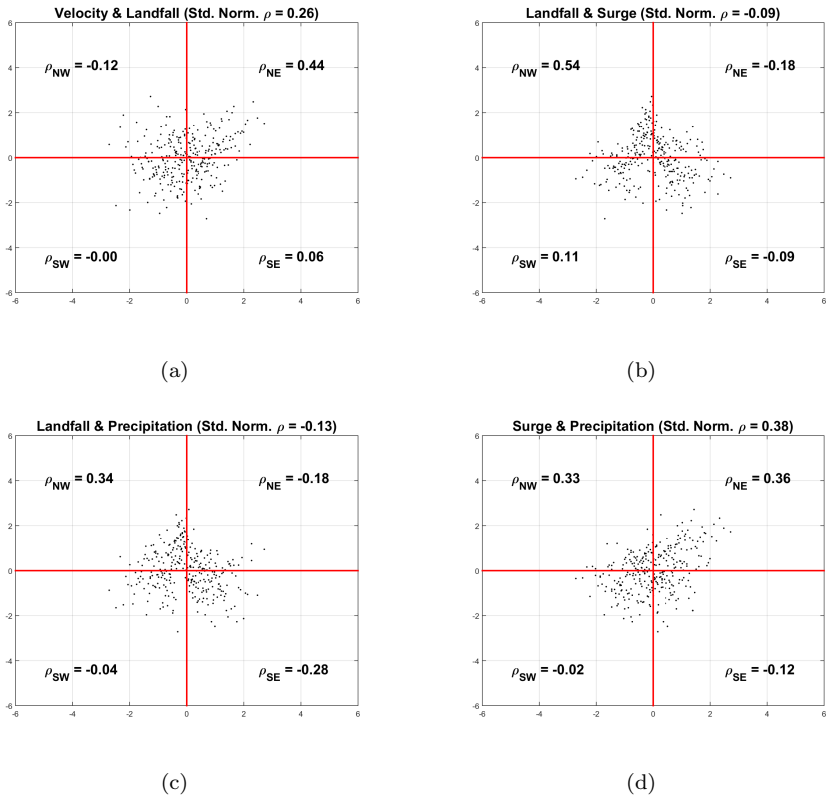

Figure A.9: Graphs of selected variable pairs transformed to standard normal where the difference between the CM_n Statistic for the Gaussian and Gumbel copulas is larger than 0.08. The correlation coefficients for the whole sample and the semi-correlations for each quadrant are given.

be valid [18].

Based on this test, we find that the Gaussian copula is a good assumption for the network presented in this study. The DER was within the 90% confidence bound of the DNR for when up to 1500 samples are drawn and the DNR was within the 90% confidence bound of the DBN for a sample size of about 600. This indicates that the joint normal copula is an adequate assumption for the BN in this study.

d-Calibration Score. Because it is possible that the DNR and DER can be equal even for a situation in which the two matrices are not equivalent, a new score measuring the validation of the network has been presented by [42] which measures the closeness between two correlation matrices using a "d-calibration"

score. The score is 1 if the matrices are equal and 0 if one matrix contains perfectly correlated variables and the other does not. The closer the score is to 1, the closer the two matrices are to each other.

The d-calibration score is calculated as follows [42]:

$$d(\Sigma_1, \Sigma_2) = 1 - \sqrt{1 - \eta(\Sigma_1, \Sigma_2)} \tag{A.11}$$

$$\eta(\Sigma_1, \Sigma_2) = \frac{|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{1/2}} \tag{A.12}$$

where Σ_1 and Σ_2 are the correlation matrices. In the case of our network, the distance between the empirical and empirical normal rank correlation matrices under the normal copula is 0.9526. This within the uncertainty bounds (90%) for a sample of about 2000. The distance between the empirical normal and normal rank correlation matrices under the normal copula is 0.9088. This is within the uncertainty bounds for a sample of about 700. This confirms that the BN construction is valid for our data set.

Appendix A.4. Conditionalizing the Network

To demonstrate the flexibility of the model, we present two states of the NPBN: (1), where the landfall location and windspeed are conditionalized to generate probable estimates for surge and precipitation and (2), where surge is conditionalized to lend insight into the hurricane characteristics that lead to high surge at Galveston Pier 21. Figure A.10 shows the BN conditionalized for angle of approach and windspeed. In this network, we have chosen to model a strong Category 3 (windspeed = 110 kts) hurricane making landfall perpendicular to the coast (angle=235 degrees). This provides interesting information about landfall location as we see that these types of hurricanes are more likely to make landfall in the central portion of the Gulf Coast (near the Upper Texas Coast Coast or Western Louisiana Coast). As expected, surge increases with respect to the base case (Figure 5).

Similarly, we can derive valuable information about the types of hurricanes that could produce high storm surge at Galveston Pier 21. For example, conditionalizing the NPBN for storm surge of 4.2 meters, yields windspeeds of 106 kts, or a medium Category 3 hurricane (Figure A.11). Furthermore, we see

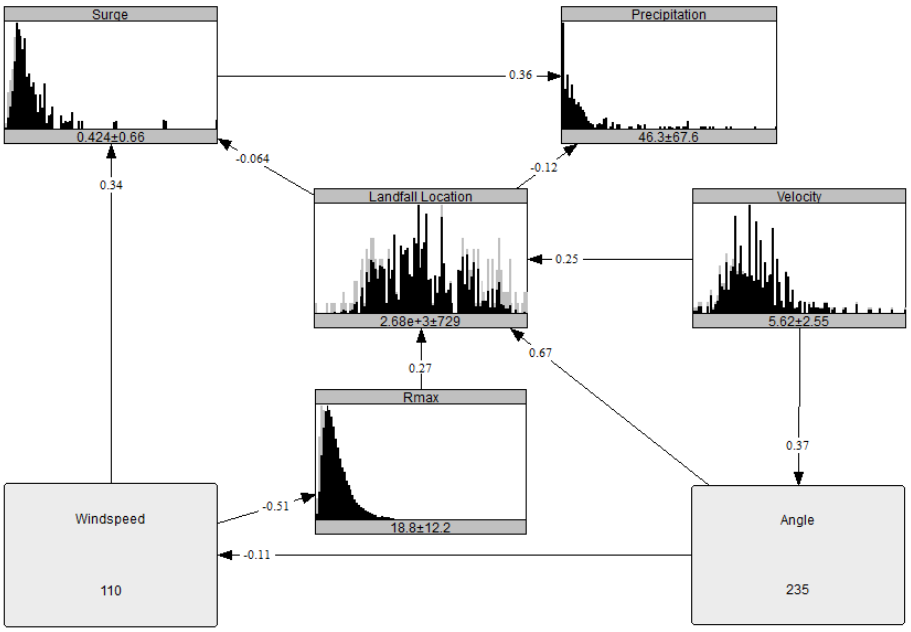

Figure A.10: Example conditionalized network for windspeed and landfall location.

that the predicted precipitation corresponding with a 4.2 meter surge will be approximately 170 mm which is a significant increase with respect to the base case.

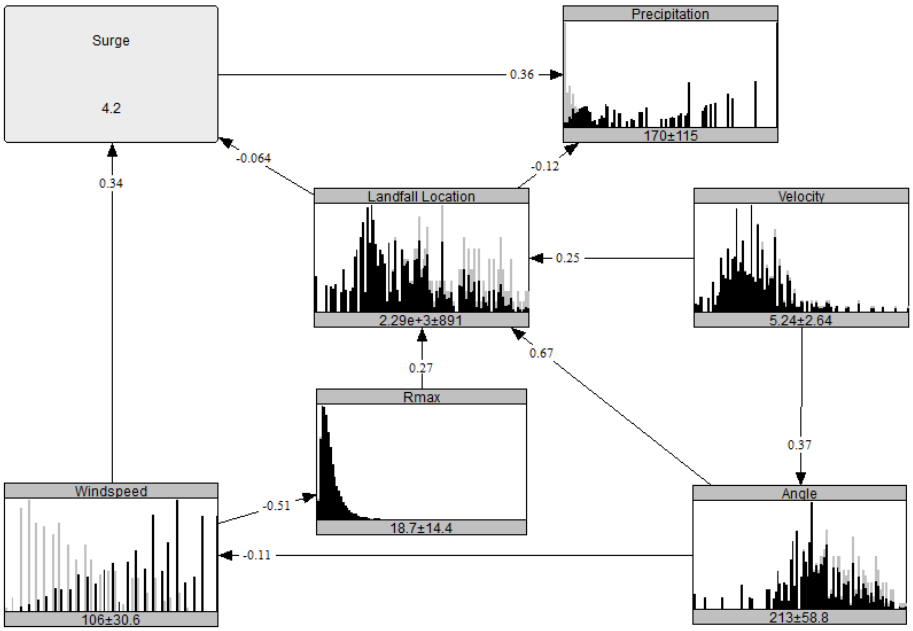

Figure A.11: Example conditionalized network for surge.