Applied Joint Probabilistic Modeling of Compound Coastal Flood Hazard: An Extension of the Joint Probability Method with Optimal Sampling

Nathan B. Geldner
PhD Candidate, School of Industrial Engineering, Purdue University, West Lafayette, USA

David R. Johnson
Professor, School of Industrial Engineering, Purdue University, West Lafayette, USA

Gabriele Villarini
Professor, The University of Iowa, Department of Civil and Environmental Engineering, Iowa City, USA

Brendan Yuil
Hydrologist, U.S. Army Corps of Engineers New Orleans District, New Orleans, USA

Angshuman Saharia
Research Scientist, The Water Institute of the Gulf, Baton Rouge, USA

Lauren Grimley
Research Scientist, The Water Institute of the Gulf, Baton Rouge, USA

Nathan Young
Senior Research Scientist, The Water Institute of the Gulf, Baton Rouge, USA

Myles McManus
Research Scientist, The Water Institute of the Gulf, Baton Rouge, USA

Hugh Roberts
Senior Vice President and Chief Operating Officer, The Water Institute of the Gulf, Baton Rouge, USA

Shubhra Misra
Senior Research Scientist, The Water Institute of the Gulf, Baton Rouge, USA

ABSTRACT: Compound coastal flooding i.e. coastal flooding driven by storm surge, rainfall, and riverine dynamics poses a significant and complex hazard. We present a novel framework for statistical modeling of this hazard as applied in a preliminary pilot study in Louisiana. This framework extends the Joint Probability Modeling with Optimal Sampling (JPM-OS), previously used for purely surge and wave driven flooding, with a stochastic rainfall field generator to produce an empirical distribution of compound surge-rainfall events with pre-computed surge and wave behavior modeled via ADCIRC + SWAN and hydrologic behavior modeled via HEC-HMS. A clustering-based discretization scheme is then applied to the sampling distribution in order to reduce set of outcomes to a size which can be tractably simulated via HEC-RAS while minimizing the square error induced by discretization. While model improvements are ongoing, the clustering-based discretization scheme is highly generalizable, provides guaranteed convergence to local optima, and performs well in preliminary analysis.

1. INTRODUCTION

Compound coastal flooding, i.e., flooding driven by interacting pluvial, riverine, and coastal dynamics, poses a significant hazard which in some areas is much greater than can be attributed to inland or coastal dynamics separately [1]–[4]. Characterizing this hazard requires two major model components: a physically driven simulation model or metamodel thereof which estimates flood depths resulting from a given storm event, and a statistical model which estimates the probability distribution of the number and characteristics of storm events in a given year. While physically driven simulation of compound flood events represents an active area of research [5], the methods discussed here focus on the statistical modeling of compound flood hazard from tropical storms specifically.

1.1. The structure of statistical models of compound tropical flood hazard

Statistical models of compound flooding from tropical cyclones consist of three major components. The first component, the recurrence rate analysis, estimates the rate at which tropical cyclones occur. The second describes the continuous joint distribution of tropical cyclone features which drive hazard when tropical cyclones occur. The third discretizes the continuous distribution of tropical cyclone features to a set of events which can tractably be run through physically driven simulations.

The recurrence rate analysis is typically done with the capture zone or kernel function weighting approach [6], [7]. The capture zone approach simply counts and averages the number of storms passing through a specified area per year. The kernel function weighting approach applies a smoothing kernel to the travel paths or “tracks” of historical storms and integrates the resulting kernel frequency density over a length of idealized coastline or region of interest.

The continuous joint distribution of tropical cyclone features is typically captures using copulas, physically driven Monte Carlo ensembles, and joint probability methods. Copulas are common in the literature and easy to use, as they require only specified marginal distributions and simplified dependence structures between hazard drivers such as peak surge and total rainfall [8]–[10], but in practice can only produce dependency structures which match one or at most two dependency measures of the true joint distribution [11], e.g. the meta-gaussian copula which captures rank correlation only. Physically driven Monte Carlo ensembles are less common, and are generated by randomly seeding tropical cyclone vortices and evolving them with deterministic meteorological simulations [12], which carries the advantages for the physical realism of individual cyclones but may or may not reflect the true joint variance structure of storm features. Joint probability methods are uncommon in compound flood hazard analysis and more commonly used for purely coastal i.e., surge and wave driven flood hazard characterization. Joint probability methods leverage empirically derived statistical relationships and conditional independence structures permitting analysts to flexibly express the joint distribution of tropical cyclone features as a series of conditional distributions or Bayesian factorization [13]–[15].

Continuous joint distributions of tropical cyclone features are typically discretized in one of three ways: naïve Monte Carlo sampling [16], structured samples [15], and optimization-driven subsampling of larger Monte Carlo or structured samples[13], [17]. The idealized discrete storm events in the resulting distribution are referred to as synthetic storms. Naïve Monte Carlo sampling directly samples from the continuous joint distribution but requires a large sample size. A structured sample can more efficiently span tropical cyclone parameter space but relies on a heuristic integration scheme to assign probability masses and may also require a large sample size. Optimization-driven subsampling is often used to reduce the set of synthetic storms to a size for which flood depths can be more tractably simulated in coastal flood risk analysis [18], [19], but requires initial simulation of the original set

[17] or unrealistic assumptions about the variance structure of conditional flood depth exceedance probabilities for Bayesian quadrature [18].

1.2. Project Context

The methods presented here were developed in the course of a preliminary pilot analysis for the Louisiana Watershed Initiative and applied in an illustrative case study to the Amite River Basin. Efforts towards a revision of this pilot study to finalize methods before eventual coastwide implementation are ongoing and will be noted as appropriate.

2. METHODS

It was decided at the outset of the project that the statistical model of joint flood hazard would extend the CLARA model used in Louisiana's 2023 Coastal Master Plan [17]. This version of CLARA used a one-dimensional capture zone (i.e. line-crossing) approach for recurrence analysis, although ongoing development has replaced this with a kernel density weighting approach. CLARA uses a joint probability method to characterize the continuous joint distribution of five tropical storm parameters at landfall: landfall location, central pressure, radius of maximum windspeed, heading angle, and forward velocity. While CLARA supports discretization via a structured set of 645 synthetic storms, it was decided due to computational constraints to use a reduced set of 50 synthetic storms using subsampling methods previously applied in the 2023 Coastal Master Plan [17] which performs well in approximating the distribution of surge hazard. In future analysis, any subsetting of the larger synthetic storm set will be conform to optimal sampling methods for compound hazard described below.

The distribution of rainfall conditionally on the five tropical cyclone parameters used in CLARA was modelled using the stochastic rainfall generator for tropical cyclone produced by Villarini et al. [20]. This generator estimates the expected rainfall associated with a synthetic storm and samples from a parameterized model of the residual variance. In doing so it captures and samples from the aleatory uncertainty in rainfall from associated with each synthetic storm. Further investigation revealed bias in the stage IV data initially used to calibrate the generator, taken from the National Centers for Environmental Prediction. This resulted in five equiprobable bias correction factors applied to rainfall fields produced by the generator, and future analysis will instead use a similar generator calibrated using the alternative Analysis of Record for Calibration dataset [21]. Additionally, it was found that the distribution of antecedent conditions could be reasonably represented using three equiprobable cases, although future analysis will instead utilize five probability-weighted cases. The joint distribution of coastal and inland flood drivers was therefore characterized with a discrete distribution of 50 probability-weighted synthetic storms each of which with an arbitrarily large set of equiprobable stochastic rainfall fields (50 as implemented although more will be used in the future), each with five bias correction factors and three antecedent conditions cases. In practice this resulted in an empirical distribution characterized by a set of 37,500 events

Flood depths for discrete events were simulated via HEC-RAS with upstream hydrological boundary conditions from HEC-HMS and downstream boundary conditions from ADCIRC+SWAN, although future analysis will utilize updated HEC-RAS with rain-on-grid for the full model domain in lieu of upstream hydrological modeling. While future analysis will use a substantially better-optimized HEC-RAS model, the computational efficiency of the HEC-RAS model available was such that the initial pilot analysis was limited to 200 HEC-RAS simulations.

This required the use of a novel optimal sampling discretization procedure for compound coastal flooding. The available budget of 200 HEC-RAS simulations was insufficient to evaluate even a single rainfall field over all five bias corrections for each synthetic storm and antecedent conditions case combination. However, ADCIRC+SWAN simulation output was already available for each synthetic storm from the 20203 Coastal Master Plan, and the HEC-HMS model used to calculate upstream hydrological boundary conditions was orders of magnitude faster than the HEC-RAS model. We therefore evaluated discharge behavior of each of 50 stochastic rainfall fields for each of the 50 synthetic storms with each of the 5 bias correction factors for each of the three antecedent conditions cases, extracted features of each simulation run which combined with features of surge behavior were taken to represent the joint distribution of compound flood drivers given the occurrence of a tropical cyclone. This distribution was then discretized using a clustering-based approach to minimize the integrated square error induced by discretization. We refer to the methods used in their totality as the extended joint probability method with optimal sampling (EJPM-OS).

2.1. Optimal sampling discretization for compound coastal flood risk

The goal of optimal sampling discretization for compound coastal flood risk is to discretize a continuous random variable or to more coarsely discretize a discrete random variable while introducing as little error as possible. We define the error induced by discretization as a loss function in Equation 1.

$$L_X(X') = \int_{\Omega} \left|\left|X(\omega) - X'(\omega)\right|\right|^2 dp(\omega) \qquad (1)$$

Here X is the original (multivariate) random variable, X' is the discretized random variable, $L_X(X')$ is the loss function or error induced by approximating X as X'. The right-hand side of the equation invokes the measure-theoretic definition of a random variable. A random variable is defined as a function $X: \Omega \rightarrow \mathbb{R}$ where Ω is a sample space consisting of possible events. Our multivariate random variables are vectors of univariate random variables $X = (X_1, X_2, \ldots)$, $X' = (X'_1, X'_2, \ldots)$. The loss function or approximation error can be interpreted as the Euclidean distance between the true representation of an event $\omega \in \Omega$, $X(\omega) \in \mathbb{R}^n$, and its approximated representation after discretization $X'(\omega) \in \mathbb{R}^n$, integrated over the space of events Ω with probability measure p.

In the case of a continuous original random variable, we approximate the continuous random variable with one constructed from an arbitrarily large random sample, resulting in Equation 2. Note that in the case that the original random variable is already discrete Equation 2 holds with equality.

$$L_X(X') \approx \sum_{\Omega} ||X(\omega) - X'(\omega)||^2 p(\omega) \qquad (2)$$

We wish to select X' so as to minimize the approximated loss function. We see that this is achieved by performing weighted k-means clustering of $X(\omega)$ and setting $X'(\omega)$ equal to the centroid of the cluster containing $X(\omega)$. This follows from equation 2. In equation 3 were-express equation 2 in terms of outcomes $x = X(\omega)$ and set $X'(\omega) = \mu_i$ where i is selected such that $X(\omega) \in S_i$ where S_i is the cluster containing $X(\omega)$, and we observe that the approximated value of our loss function from equation 2 is exactly equal to the within-cluster variance which is minimized by observation-weighted k-means clustering as we see in equation 3.

$$L_X(X') \approx \sum_{i=1}^{k} \sum_{x \in S_i} \left|\left|x - \mu_i\right|\right|^2 p(x) \qquad (3)$$

Note that k is the number of clusters or discrete values of X', which we set a-prior based on our computational constraints. While k-means clustering algorithms guarantee convergence only to locally optimal clusterings, repeated optimization with randomly initialized centroids ensures results which are close to globally optimal.

The most significant weakness of this approach as initially implemented is that optimal sampling on boundary condition features i.e. surge and discharge information is not necessarily the same as optimal sampling on peak water surface elevation, which is ultimately the hazard of interest. However, HEC-RAS is a deterministic simulation, so a discretization of the hazard distribution which induces no error in the distribution of boundary conditions would similarly induce no error in the distribution peak water surface elevations. While a discretization which induces no error is of course impossible both for fundamental reasons and because the features of surge and discharge behavior used for discretization are much lower-dimensional than the full spatially explicit time series used as boundary conditions, this leads us to believe that a discretization which performs well in minimizing error in the distribution of boundary conditions will similarly perform well in minimizing the error in peak water surface elevation. Future analysis using a much better-optimized HEC-RAS model will permit us to empirically investigate this assumption. Additionally, due to the availability of better-optimized HEC-RAS models, future work will use the output of HEC-RAS model runs with a coarser computational mesh for discretization rather than HEC-HMS and ADCIRC+SWAN results.

An additional limitation of this method in practice is that there is no observation at the exact centroid of each cluster, so the observation nearest each cluster centroid is used instead. Investigations into how many stochastic rainfall fields per synthetic tropical cyclone are required to adequately characterize the aleatory uncertainty in rainfall, as well as how many clusters are required to adequately capture the variability of the full distribution are ongoing. Work is also ongoing to refine pre-processing steps and implementation details of the clustering-based optimal sampling scheme, detailed below.

2.2. Implementation of optimal sampling discretization

We start by extracting peak discharge, runup time, and drawdown time from each major inlet to the HEC-RAS domain, as well lag time between time of peak surge and peak discharge and surge characteristics including average peak surge depth, runup time, and drawdown time among representative points. Discharge runup times are calculated by treating the discharge from the time at which discharge first exceeds its mean value over the hydrograph up to the time of peak discharge as the left half of a gaussian density function and calculating the corresponding standard deviation. Drawdown times were similarly calculated from the time of peak discharge up to the point at which discharge receded below its mean value over the hydrograph. Surge runup and drawdown times were similarly calculated. A log transformation was applied to peak, runup, and drawdown of discharge and surge in the preliminary study due to pronounced skewness, but future analysis will not apply the log transformation as doing so reduces the effective weight of extreme results. All features were then standardized to have mean zero and standard deviation equal to 1, and future analysis will have all feature scaled by the square root of an assigned importance weight. The importance weights will likely be assigned such that the total weight assigned to surge behavior is equal to that of discharge behavior, and half of the weight placed on both surge and discharge behavior will be placed on peak values. Preliminary results are believed to have placed insufficient weight on surge compared to discharge, and insufficient weight on peak values compared to runup and drawdown rates. Further analysis is required to investigate the impacts of these feature weights. Events were heuristically observation-weighted according to the CLARA-derived probability mass of their corresponding synthetic storms by use of repeated observations, permitting us to treat the set of events as a random sample of equiprobable events. Future analysis will permit continuous observation weights instead.

Following extraction and standardization (and feature weighting in future analysis), we apply principal component analysis to the empirical distribution. The dimensionality and size of the sample did not require dimensionality reduction in preliminary analysis, but principal component analysis was helpful in holistically evaluating the performance of the sampling approach. Future analysis may or may not require dropping small principal components for computational tractability.

From this point the sample was clustered and discretized such that the observation nearest the centroid of each cluster was assigned the summed probability mass of observations in the respective cluster. Several synthetic storms were unrepresented in the resulting discretization, likely due to aforementioned under-weighting of storm surge features. In the preliminary analysis this led to an adjustment which replaced certain cluster centroids with nearby events from unrepresented synthetic storms so as to minimize additional error induced by the adjustment, but this adjustment is unlikely to be included in future analyses both due to the anticipated effects of feature-weighting prior to clustering on the diversity of synthetic storms in the optimal sample and due to the larger set of synthetic storms which will be used in future analysis.

3. RESULTS

While the methods described above as implemented in the preliminary analysis reflect a non-negligible contribution to the state of practice of statistical modeling of compound coastal flood hazard, they reflect development for an exploratory and preliminary analysis and are not fully reflective of the final methods which will be used for the Louisiana Watershed Initiative or the revised pilot study of the Amite River Basin. As noted throughout, it contains several methodological details which will be revised and improved upon. For this reason, this report does not contain any results or figures which could be construed as flood maps or hazard estimates. Instead, the results presented here reflect the performance of the optimal sampling discretization scheme.

3.1. Fidelity of optimal sampling results to original sample

The clustering-based optimal sampling scheme for compound coastal flood hazard presented here, despite the various implementation issues described which had yet to be adjusted, performed surprisingly well at approximating a distribution characterized by 37,500 events in 16 dimensions (50 synthetic storms, 50 rainfall fields, five bias correction factors, and three antecedent conditions cases with peak surge, surge runup and drawdown, lag time between peak surge and peak discharge, and peak discharge and discharge runup and drawdown for four inlets). Figure Figure 1: Cumulative distribution functions of first principal component value and peak discharge at the largest inlet of the HEC-RAS domain, characterized by the original sample and the optimal subsample. shows the cumulative distribution functions of the first principal component of the sample and of peak discharge at the largest inlet to the HEC-RAS domain. The optimal subsample matches the distribution of the first principal component of the original sample almost exactly. The optimal subsample appears to underestimate peak discharge for extreme events. We expect this to improve in future analyses when we no longer apply the log transformation in pre-processing.

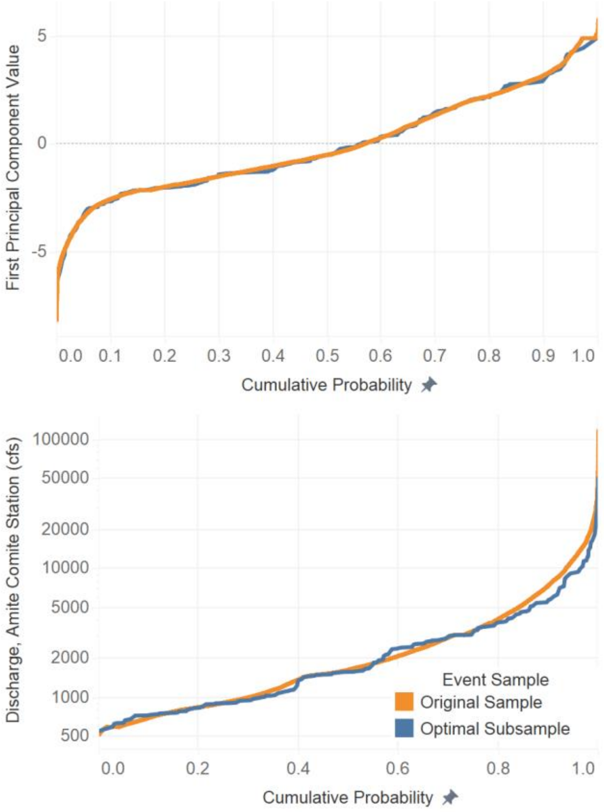

Figure 1: Cumulative distribution functions of first principal component value and peak discharge at the

largest inlet of the HEC-RAS domain, characterized by the original sample and the optimal subsample.

3.2. Agreement between coastal hazard characterization of original sample and optimal subsample

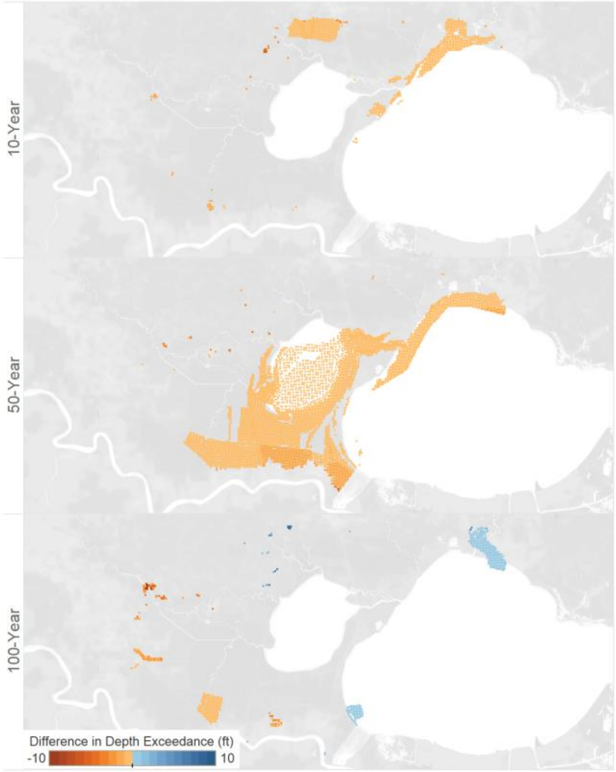

Figure 2: Difference between surge-only flood hazard estimated by EJPM-OS as implemented and previous CLARA methods, expressed in feet at the 10-, 50-, and 100-year return periods. Only pixels with a difference of 6 inches or greater are shaded.

The clustering-based optimal subsampling approach will typically result in clusters containing storm events generated from more than one synthetic storm and therefore results in changes to the probability masses assigned to each synthetic storm compared to those in the original sample. To investigate the magnitude of this effect, we compared surge-driven flood hazard estimated by the original CLARA model with surge-driven flood hazard estimated using probability masses for each synthetic storm corresponding to the results of the optimal subsampling procedure from EJPM-OS. The results are shown in Figure 2. We see broad agreement between the methods at the 10- and 100-year return periods, although we do see that the EJPM-OS-derived probability masses underestimate the 50-year flood depth by about a foot in a section of the model domain. We expect this performance to only improve as we increase the weight placed on surge features in the optimal sampling process.

4. CONCLUSIONS

The extended joint probability method with optimal sampling (EJPM-OS) represents a novel approach for statistical modeling of compound coastal flood hazard. Preliminary implementation has shown good performance over several measures, and more detailed evaluation of assumptions and methodological details will be published in the future as a part of the Louisiana Watershed Initiative.

The clustering-based optimal sampling procedure used in EJPM-OS is highly generalizable to statistical characterization of natural hazards where the outcome of interest of a random event is calculated with a computationally expensive model, and for which boundary conditions or lower-fidelity estimates (via coarser model structure or metamodeling) can be produced more efficiently. It can be applied directly in cases where the distribution of events is represented as an empirical distribution or random sample, and it can be applied to a large Monte Carlo sample of an arbitrary joint distribution without requiring any assumptions about the variance structure of the hazard. The optimality guarantee associated with this approach, that of minimizing integrated square error in discretization, is highly appropriate from the perspective of viewing natural hazards through the lens of multivariate random processes.

5. ACKNOWLEDGEMENTS

We acknowledge the support and guidance of Louisiana state agency personnel from the Council of Watershed Management supporting the Louisiana Watershed Initiative (LWI), as well as the LWI Data and Modeling Transition Zone Technical Advisory Group

6. REFERENCES

[1]J. Zscheischler *et al.*, “Future climate risk from compound events,” *Nat. Clim. Change*, vol. 8, no. 6, pp. 469–477, 2018.

[2]H. Moftakhari, D. F. Muñoz, J. Y. Song, A. Alipour, and H. Moradkhani, “Challenges for Appropriate Characterization of Compound Coastal Hazards,” in *Geo-Extreme 2021*, 2021, pp. 58–68.

[3]A. AghaKouchak *et al.*, “Climate extremes and compound hazards in a warming world,” *Annu. Rev. Earth Planet. Sci.*, vol. 48, pp. 519–548, 2020.

[4]A. Sebastian, “Compound flooding,” in *Coastal Flood Risk Reduction*, Elsevier, 2022, pp. 77–88.

[5]F. L. Santiago-Collazo, M. V. Bilskie, and S. C. Hagen, “A comprehensive review of compound inundation models in low-gradient coastal watersheds,” *Environ. Model. Softw.*, vol. 119, pp. 166–181, 2019.

[6]N. C. Nadal-Caraballo, V. M. Gonzalez, and L. Chouinard, “Storm Recurrence Rate Models for Tropical Cyclones: Report 1,” ENGINEER RESEARCH AND DEVELOPMENT CENTER VICKSBURG MSMCGILL UNIV MONTREAL …, 2019.

[7]M. Bensi and T. Weaver, “Evaluation of tropical cyclone recurrence rate: factors contributing to epistemic uncertainty,” *Nat. Hazards*, vol. 103, no. 3, pp. 3011–3041, 2020.

[8]H. Kim, G. Villarini, R. Jane, T. Wahl, S. Misra, and A. Michalek, “On the generation of high-resolution probabilistic design events capturing the joint occurrence of rainfall and storm surge in coastal basins,” *Int. J. Climatol.*, 2022.

[9]A. Couasnon, A. Sebastian, and O. Morales-Nápoles, “A copula-based Bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment scale: An application to the Houston Ship Channel, Texas,” *Water*, vol. 10, no. 9, p. 1190, 2018.

[10] Z. Hao and V. P. Singh, “Compound events under global warming: a dependence perspective,” *J. Hydrol. Eng.*, vol. 25, no. 9, p. 03120001, 2020.

[11] M. Hofert, I. Kojadinovic, M. Maechler, J. Yan, J. G. Nešlehová (evTestK()), and R. M. (fitCopula ml(): code for free mixCopula weight parameters), “copula: Multivariate Dependence with Copulas.” Jan. 25, 2023. Accessed: Feb. 08, 2023. [Online]. Available: https://CRAN.R-project.org/package=copula

[12] A. Gori, N. Lin, and D. Xi, “Tropical cyclone compound flood hazard assessment: From investigating drivers to quantifying extreme water levels,” *Earths Future*, vol. 8, no. 12, p. e2020EF001660, 2020.

[13] A. Gori and N. Lin, “Projecting compound flood hazard under climate change with physical models and joint probability methods,” *Earths Future*, p. e2022EF003097, 2022.

[14] G. R. Toro, D. T. Resio, D. Divoky, A. W. Niedoroda, and C. Reed, “Efficient joint-probability methods for hurricane surge frequency analysis,” *Ocean Eng.*, vol. 37, no. 1, pp. 125–134, 2010.

[15] D. R. Johnson, J. R. Fischbach, and D. S. Ortiz, “Estimating surge-based flood risk with the coastal Louisiana risk assessment model,” *J. Coast. Res.*, no. 67 (10067), pp. 109–126, 2013.

[16] Y. Peng, K. Chen, H. Yan, and X. Yu, “Improving flood-risk analysis for confluence flooding control downstream using Copula Monte Carlo method,” *J. Hydrol. Eng.*, vol. 22, no. 8, p. 04017018, 2017.

[17] J. R. Fischbach, D. R. Johnson, M. T. Wilson, N. B. Geldner, and C. Stelzner, “Draft Coastal Master Plan: Risk Assessment Model Improvements,” *Version I*, pp. 1–78, 2023.

[18] G. R. Toro, A. W. Niedoroda, C. W. Reed, and D. Divoky, “Quadrature-based approach for the efficient evaluation of surge hazard,” *Ocean Eng.*, vol. 37, no. 1, pp. 114–124, 2010.

[19] K. Yin, S. Xu, and W. Huang, “Estimating extreme sea levels in Yangtze Estuary by quadrature joint probability optimal sampling method,” *Coast. Eng.*, vol. 140, pp. 331–341, 2018.

[20] G. Villarini, W. Zhang, P. Miller, D. R. Johnson, L. E. Grimley, and H. J. Roberts, “Probabilistic rainfall generator for tropical cyclones affecting Louisiana,” *Int. J. Climatol.*, vol. 42, no. 3, pp. 1789–1802, 2022.

[21] H. Kim and G. Villarini, “Evaluation of the Analysis of Record for Calibration (AORC) rainfall across Louisiana,” *Remote Sens.*, vol. 14, no. 14, p. 3284, 2022.