Flood Data Collection And Processing For A Machine Learning Model

Cameron Cook, Maryam Pakdehi
FAMU-FSU College Of Engineering, Tallahassee, FL

Introduction & Background

Importance of Flood Depth Estimation in Flood Risk Management

- Flooding affects tens of thousands of people directly every year globally. With the exacerbation of global warming, flooding effects are intensifying, disrupting businesses, infrastructure, and environments.

- Inaccurate predictions can result in loss of lives and property due to inadequate preliminary infrastructure investments or individual preparation.

- There are different difficulties in flood estimation depending on the method used. Estimating floods through physical models requires thorough data, and large-scale inundation models are prone to overestimation and an over-reliance on topographic data, making them unable to address most regions susceptible to flooding. In addition, training a model on a lack of diverse data can lead to inaccurate predictions for other locations.

Machine Learning in Flood Forecasting

- Machine learning has the capacity to improve the accuracy of flood estimations by considering a wide range of location-specific data, while considering location-specific influences, to create algorithms adaptable to any specific flood situation. In addition, they are incredibly cost-effective and hundreds of times quicker to run, helping determine effective protection strategies for different watersheds, and allows more local policy to be effectively implemented.

- The study developed an ML-based model, combining an Artificial Neural Network algorithm with feature selection methods and geospatial data to train, cross-verify, and test a generic model capable of hindcasting flood depths for any location. It was first evaluated and trained on Hurricane Ida's flooding in the HUC6 Lower Hudson watershed in the United States, then on other previously-researched events within the same watershed, and currently on other events and areas in the US.

Research Objective and Scope

- The primary objective of my contribution to the research is collecting and processing flood event data from nine hurricanes and six watersheds between 2011 and 2022, in the form of meteorological and geographical data input from stations throughout each affected watershed. These specific scenarios were chosen to test the ML model on varying conditions. The flood-prone, coastal nature of these watersheds, and their exposure to flooding from recent hurricanes, make them sufficient choices for analysis by the ML model. I then compared data to analyze the correlation between different data types.

Abstract

This research underscores the critical role of precise flood depth estimation in flood risk management and introduces a machine learning methodology tailored to enhance the accuracy of such estimations through hindcasting. The study primarily investigates the applicability and adaptability of a machine learning neural network model, initially developed for Hurricane Ida in the Lower Hudson watershed, by testing its performance on different hurricanes and locations. The goal is to evaluate the model's effectiveness across varying locations, conditions, and scenarios, and to understand the factors influencing the machine learning model's performance.

To train and validate the machine learning algorithms, data from the NOAA and the National Hurricane Center were collected, filtered, and analyzed from nine hurricanes' impacts between 2011 and 2022, each on one of six watersheds. These vital parameters included stream gauge water level, soil moisture, rainfall, tide, wind speed, and storm track data. This dataset was essential to train and test the machine learning model to grasp the complexities of flood dynamics and enhance its predictive capabilities.

While the results of this extensive study are pending, the anticipated outcomes aim to offer valuable insights into the viable efficacy of machine learning in predicting flood depths. This research is poised to contribute significantly to the field of flood risk management, highlighting the potential of advanced predictive models in ensuring community safety and resilience against the backdrop of increasingly frequent and severe rainfall events.

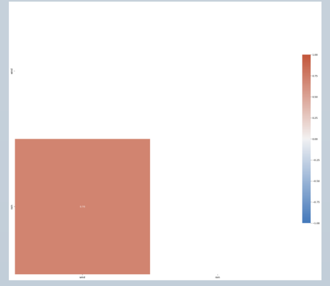
Methods

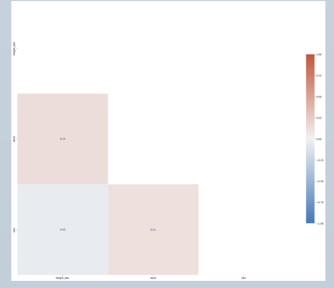
- Various hydrological, geographical, and meteorological datasets were acquired for the duration each researched hurricane produced rainfall in each system's investigated watershed. Stream gauge data was collected to determine water level in rivers, tide data was collected to include tidal effects on flooding, soil moisture data was obtained to determine antecedence, high water mark (HWM) data was used as a target variable for verification of flood height, and rainfall, maximum wind speed, and storm track data were collected due to their major roles in causing flooding.

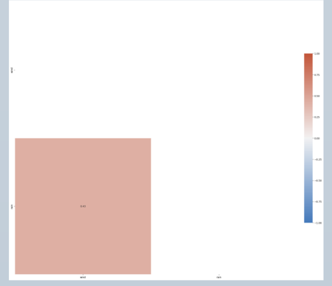
- All data with the exception of storm track data was downloaded from U.S. Government websites such as the NOAA and NHC, recorded by numerous U.S. Government-maintained sites. These were saved as Comma Separated Values files and processed through a python code filterer. Extraneous data were removed and the rest was formatted in the first round of filtration. The second round filtered the intermediary file to determine each station's most notable data point; maximums and total values for each site in a watershed during their respective storms.

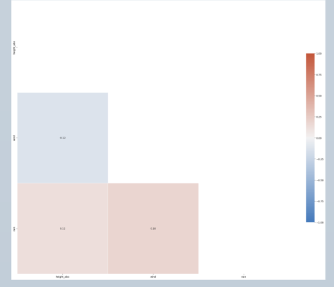
- ArcGIS Pro was utilized to display spatial interpolation of the filtered data sets to estimate conditions throughout the entire watershed that lack sites directly representing location-specific data. This helps generate logical data more completely within a watershed, and also allow manual removal of possibly erroneous data points by visualizing data.

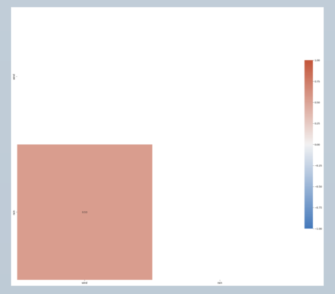
Correlation Matrixes

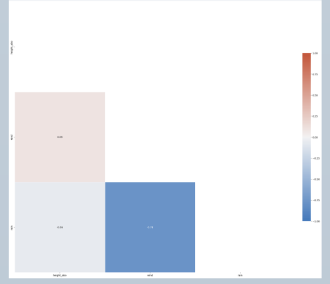

Florence

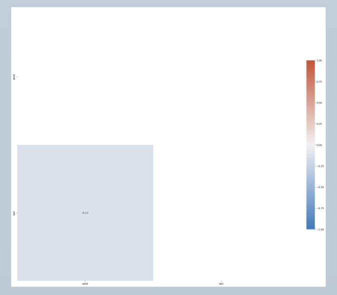

Ida


Irene

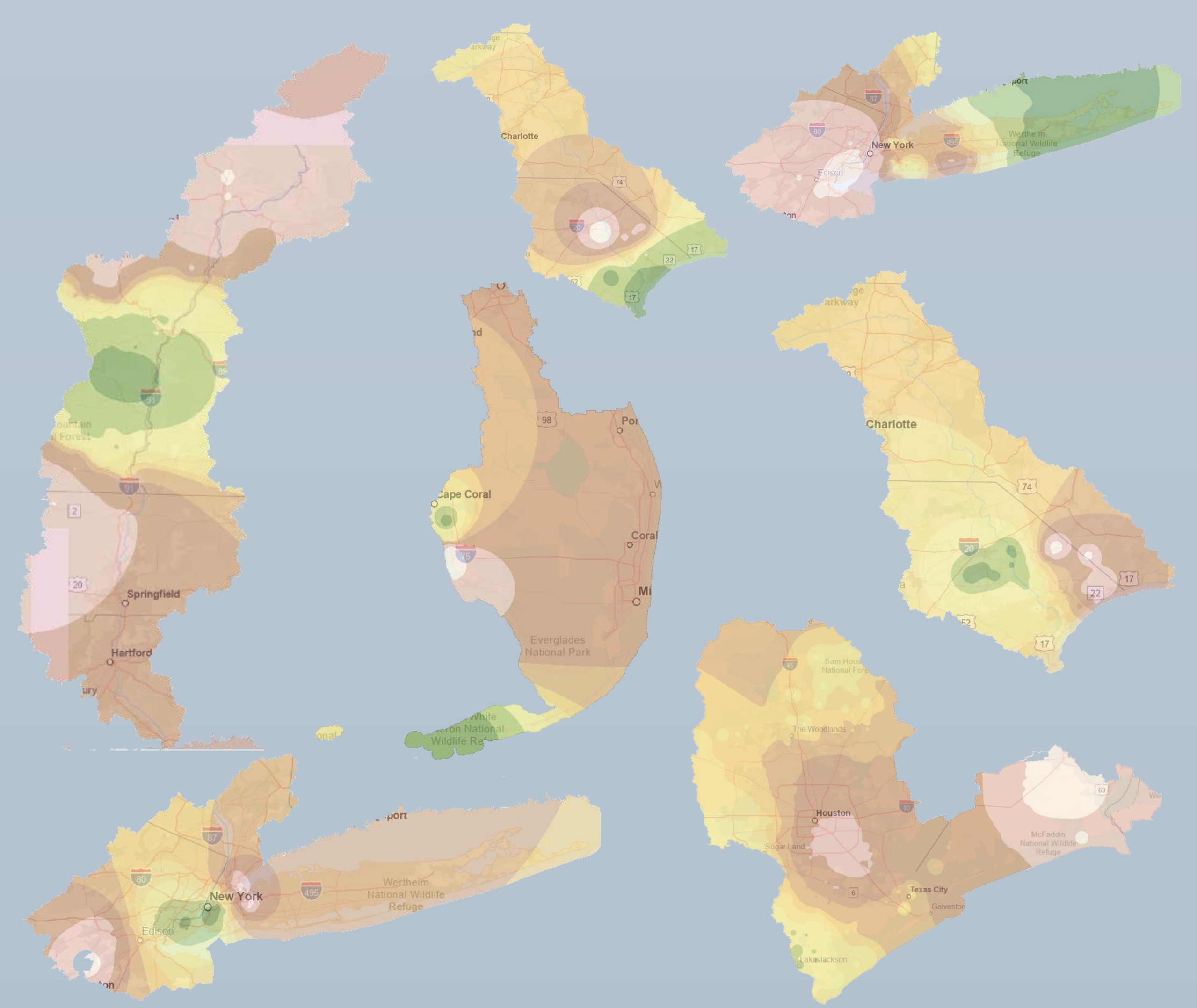

Matthew


Sandy


Irma

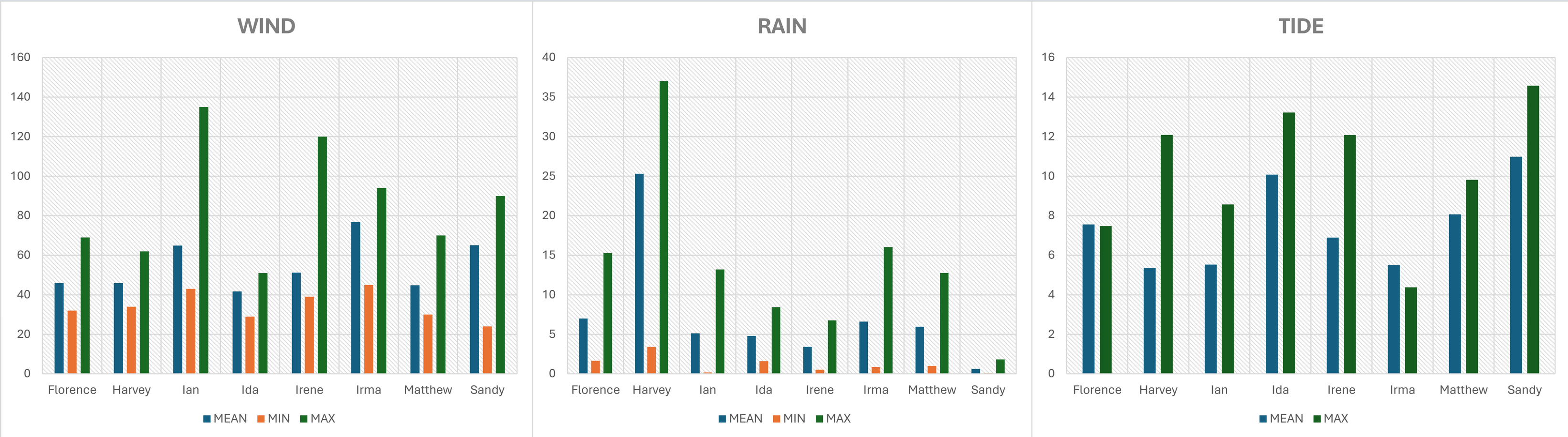

Harvey

Maps From ArcGIS Pro



Rainfall data from seven hurricanes in their respective analyzed watersheds; interpolated using ArcGIS Pro.

Table Data



Results & Discussion

Regarding the data collected, there is not a consistent correlation between all storms between rain, wind, tide, and HWM data, with the exception of occasional consistency between rain and wind, and a lack of any correlation between most other variables. The strongest correlation is found between wind and rain from Hurricane Irma (-0.78), which result in the removal of one variable from further analysis in this scenario to avoid problems with the machine learning model. This data implies significance of location and situation-specific variables in flood hindcasting.

The lack of correlation between wind, rain, tide, and HWM show that any single variable can be responsible for a storm's destructive force. Testing the viability of the machine learning model in other regions and other times is important to determine the model's versatility for the sake of more accurate localized forecasts, as results will indicate how well this and similar models can be applied to flood management forecasts, and in turn, area-specific protection methods.

References

Antwi-Agyakwa, K. T., Afenyo, M. K., & Angnuureng, D. B. (2023). Know to Predict, Forecast to Warn: A Review of Flood Risk Prediction Tools. *Water*, *15*(3), 427. https://doi.org/10.3390/w15030427

Apan, A., Keogh, D., King, D., Thomas, M., Mushtaq, S., & Baddiley, P. (2010). *The 2008 floods in Queensland: A case study of vulnerability, resilience and adaptive capacity Case studies of extreme events*.https://nccarf.edu.au/wp-content/uploads/2019/03/Apan_2010_The_2008_Floods_in_Queensland.pdf

Bergillos, R. J., Rodriguez-Delgado, C., & Iglesias, G. (2019). Wave farm impacts on coastal flooding under sea-level rise: A case study in southern Spain. *Science of the Total Environment*, *653*, 1522–1531.https://doi.org/10.1016/j.scitotenv.2018.10.422

Clark, A. (n.d.). *THE GROWING THREAT OF URBAN FLOODING: 2018*. Retrieved March 2, 2024, from https://www.tamug.edu/newsroom/2018articles/UrbanFloodingReport.pdf

Data Access | National Centers for Environmental Information (NCEI). (n.d.). Www.ncei.noaa.gov. Retrieved March 2, 2024, from https://www.ncei.noaa.gov/access/search/data-search/local-climatological-data

Find My Organization | EBSCO. (n.d.). Www.ebsco.com. Retrieved March 2, 2024, from https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=e000xna&AN=170628&site=ehost-live.

Khosravi, K., Shahabi, H., Pham, B. T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly, H.-B., Gróf, G., Ho, H. L., Hong, H., Chapi, K., & Prakash, I. (2019). A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods. *Journal of Hydrology*, *573*, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073

Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "Assessment of economic flood damage." *Natural Hazards and Earth System Science*, *10*(8), 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010

National Hurricane Center. (2024). *National Hurricane Center*. Noaa.gov. https://nhc.noaa.gov

Neave Interactive. (2017). *Zoom Earth - Explore satellite and aerial images of the Earth*. https://zoom.earth/

Pakdehi, M., Ahmadisharaf, E., Nazari, B., & Cho, E. (2023). Transferability of machine learning-based modeling frameworks across flood events for hindcasting maximum river flood depths in coastal watersheds. *Natural Hazards and Earth System Sciences Discussions*, 1–57. https://doi.org/10.5194/nhess-2023-152

Sign in - Google Accounts. (n.d.). Accounts.google.com. https://code.earthengine.google.com/

Station Selection-NOAA Tides & Currents. https://tidesandcurrents.noaa.gov/stations.html?type=Water+Levels

Tropical Cyclone Radar Loops. Tropical Cyclone Radar Loops. https://bmcnoldy.earth.miami.edu/tropics/radar/

USGS Surface-Water Historical Instantaneous Data for the Nation: Build Time Series. (n.d.). Waterdata.usgs.gov. https://waterdata.usgs.gov/nwis/uv/?referred_module=sw

Xian, S., Yin, J., Lin, N., & Oppenheimer, M. (2018). Influence of risk factors and past events on flood resilience in coastal megacities: Comparative analysis of NYC and Shanghai. *Science of the Total Environment*, *610-611*, 1251–1261. https://doi.org/10.1016/j.scitotenv.2017.07.229

Acknowledgements

Research Mentor – Maryam Pakdehi; assisted with poster presentation, research, research materials and resources, and machine learning methodology. Created machine learning model used in this research. Thank you so much for allowing me to assist in your research, and for helping me explore machine learning as well.

UROP Leaders – Remi Ventura and Carolyn Emerson assisted with ensuring I persevered in the early stages of finding a research project, and made progress throughout assisting my mentor. Thanks to them for helping me throughout my UROP experience.