UNCERTAINTY-AWARE DEEP LEARNING FRAMEWORK FOR FORECASTING COASTAL WATER LEVEL IN VIRGINIA BEACH

Mahmudul Hasan, Malachi Schram & Diana McSpadden
Department of Data Science, Jefferson Lab
Newport News, VA, USA
{mdhasan, schram, dianam}@jlab.org

Sridhar Katragadda & Alisa Udomvisawakul
City of Virginia Beach
Virginia Beach, VA, USA
{skatrag, audomvisawakul}@vbgov.com

Heather Richter & Frank Liu
Old Dominion University
Norfolk, VA, USA
{hrichter, fliu}@odu.edu

ABSTRACT

Coastal areas like Virginia Beach, USA, are increasingly vulnerable to flooding. To mitigate the impact of flooding, it is crucial for the City of Virginia Beach to have reliable 72-hour-ahead (3 days) forecasts of water levels at key gauge locations. To support this effort, several sensors have been installed throughout the city to monitor water levels and other environmental parameters such as wind speed, precipitation, and atmospheric pressure. Leveraging sensor data from one of these locations, we developed an uncertainty-aware deep learning (DL) model to forecast water levels. We employed deep quantile regression (DQR) to quantify variability in the predictions and examined the performance of three different model architectures. In addition to exclusively including historical data, we investigated the improvement wind forecasts provide to the accuracy of 72-hour-ahead water level predictions. The results show a twelvefold improvement in the flood forecast for a real flooding event.

1 INTRODUCTION

Robust and reliable predictive models are required to address the frequent and intense flood events in vulnerable coastal areas such as Norfolk and Virginia Beach, Virginia (Roy et al., 2025; Allen et al., 2023). To proactively address future hurricanes and nor'easters, the City of Virginia Beach has undertaken initiatives to enhance real-time water level monitoring through a network of Internet of Things (IoT) sensors. As part of these initiatives, the deployment of sensors began with 10 U.S. Geological Survey (USGS) tide gauges installed in 2016, later expanding to 50 additional sensors in recent years through the regional StormSense project (Loftis et al., 2018). This sensor network provides critical real-time data, enabling decision-makers to monitor adverse weather events, manage reservoir levels, and support road closures. The goal is to develop long-term flood forecasting models (three to seven days ahead) by leveraging the sensor data. Street-level flooding for 36-hour forecasts in the Hampton Roads region was initially conducted using the Tidewatch hydrodynamic model, developed by the Virginia Institute of Marine Science (Loftis et al., 2019). However, traditional hydrodynamic models require substantial computational resources for large-scale simulations (Kumar et al., 2023). These models rely on physical laws and utilize high-resolution topographic data along with well-defined boundary conditions (Teng et al., 2017; McSpadden et al., 2024). They also face challenges in capturing uncertainties and intricate nonlinear dynamics. To address these challenges, adaptive and efficient flood forecasting methods are required. Artificial intelligence (AI) and DL have emerged as powerful tools to overcome these limitations, as they can process large datasets, identify complex patterns, and adapt to changing conditions without relying on explicit physical formulations (Adikari et al., 2021; Dtissibe et al., 2024). Consequently, numerous studies in the literature have explored machine learning approaches for water level and flood

forecasting (Darabi et al., 2021; Fang et al., 2021; Wu et al., 2020; Zhang et al., 2023). Leveraging advancements in computational speed and artificial intelligence techniques, we employ a deep quantile regression model to forecast water levels three days in advance at a key gauge location in Virginia Beach, which serves as a critical indicator of flooding on the surrounding streets. To assess the model's effectiveness, we will evaluate its performance in forecasting a real flooding event that occurred in May 2022.

2 Data Collection and Preparations

We collected data (2016 to 2022) from the Little Neck Creek (LNC) location (Appendix C Figure 7) which is heavily influenced by tide. In total, eight variables are listed in the data set. They are: Date-Time, Water Level (feet), Wind Speed (mph), Wind Direction (Degree), Atmospheric Pressure (hPa), Precipitation (inches), Temperature (°F), and Dew Point (°C). Although data were collected in a six-minute frequency, due to sensor dysfunction, the data was resampled into one-hour frequency by averaging the available readings to address missing values. The initial dataset contained wind speed and wind direction (0° to 360°). An initial analysis shows weak correlations between wind speed and wind direction with water level (0.07 and -0.16 for wind speed and wind direction, respectively). Therefore, wind speed and wind direction were decomposed into their respective vector components in the X and Y directions to better capture the directional influence of wind on water level. The decomposition was carried as: $V_x = V\sin(\theta)$ and $V_y = V\cos(\theta)$. Here, V represents the wind speed, θ is the wind direction in radians, V_x is the wind X component, and V_y is the wind Y component. Following the meteorological convention, wind direction (θ) is measured clockwise from true north. After applying these transformations, the correlation coefficients of decomposed wind components with water level increased to 0.21 and 0.24, respectively (Appendix A Figure 4). These results demonstrate that wind components (Wind X and Wind Y) exhibit a stronger correlation with water level compared to the raw wind speed and direction. Next, the entire data set was scaled between 0 and 1 using the MinMax Scaler.

3 Modeling

Deep Quantile Regression (DQR) has been used to model the relationship between the predictors and the target variable. This approach estimates conditional quantiles, which make it robust against outliers in the response measurements (Koenker, 2005). This also allows the model to capture asymmetric uncertainty, which is often present in real-world data such as extreme water level events. DQR estimates the conditional quantile as: $\hat{Q}_\tau(y|x) = G_\tau(x|w)$, where $\tau \in (0, 1)$ represents the quantile level and $G_\tau(x|w)$ is the nonlinear relationship between the input x and the target variable, which is approximated by a Deep Neural Network (DNN). The DNN provides predictions for each of the predefined quantiles by minimizing the quantile loss function, defined as:

$$\mathcal{L}_\tau = \max\left(\tau \cdot (y_i - \hat{y}_i), (\tau - 1) \cdot (y_i - \hat{y}_i)\right) \tag{1}$$

Here y_i is the observed value and $\hat{y}_i$ is the predicted value for the corresponding quantile. Three different DNN architectures such as Multi-Layer Perceptron (MLP), 1D Convolutional Neural Network (Conv1D), and Long Short-Term Memory (LSTM) were utilized to implement the DQR model (Appendix B Figure 5). Different hyperparameters are selected for the three models based on hyperparameter optimization conducted using the Optuna framework Akiba et al. (2019). Key parameters such as learning rate, hidden layer size, number of layers, dropout rates, batch size and activation function were adjusted during the optimization process (Appendix B Table 3).

4 Results

A total of 48,464 data points (with a one-hour frequency) from the years 2016 to 2021 were used for training, with 20% of the training data reserved for validation. Data from 2022 were used to test the model. During training, 24-hour look-back (historical) data were used because the water level of the LNC gauge location is highly influenced by tides, exhibiting a daily cyclic pattern with two peak levels within the 24-hour period. The objective of this study is to forecast 72 hours (3 days) ahead. The

Table 1: Accuracy and Uncertainty Quantification (UQ) calibration results for the dataset at different look-ahead times.

Model	Look Ahead (hrs)	Accuracy			UQ Calibration		
		MAE ↓	RMSE ↓	R^2 ↑	MACE ↓	RMSCE ↓	MCA ↓
MLP	6	0.21	0.29	0.90	0.09	0.10	0.09
	12	0.26	0.36	0.84	0.06	0.06	0.06
	24	0.31	0.45	0.76	0.08	0.09	0.08
	48	0.43	0.60	0.55	0.03	0.04	0.04
	72	0.46	0.65	0.47	**0.02**	**0.03**	**0.02**
Conv1D	6	**0.18**	**0.24**	**0.93**	0.08	0.09	0.08
	12	**0.23**	**0.32**	**0.88**	0.06	0.07	0.07
	24	**0.29**	**0.42**	**0.79**	**0.01**	**0.02**	**0.01**
	48	**0.39**	**0.56**	**0.62**	0.04	0.04	0.04
	72	**0.43**	**0.61**	**0.51**	**0.02**	**0.03**	**0.02**
LSTM	6	0.25	0.34	0.88	**0.03**	**0.04**	**0.03**
	12	0.25	0.37	0.84	**0.04**	**0.05**	**0.04**
	24	0.30	0.45	0.75	0.02	0.03	0.02
	48	**0.39**	0.58	0.56	**0.01**	**0.02**	**0.01**
	72	**0.45**	0.66	0.46	**0.02**	**0.03**	**0.02**

model is trained for multiple look-ahead intervals, including 6, 12, 24, 48, and 72 hours. For each look-ahead, we trained the model for seven quantiles: $[0.001, 0.023, 0.159, 0.5, 0.841, 0.977, 0.999]$. In all cases, only the furthest point is forecasted.

A set of standard metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) value, were used to evaluate and compare the predictive performance of the DL models. Additionally, several metrics from the Uncertainty Toolkit Chung et al. (2021), such as Mean Absolute Calibration Error (MACE), Root Mean Square Calibration Error (RMSCE), and Miscalibration Area (MCA), are employed to assess the models' ability in uncertainty quantification (UQ). For all models, predictive performance decreases as the look-ahead period increases. Table 1 presents the values of the evaluation metrics for the three models at look-ahead intervals of 6, 12, 24, 48, and 72 hours. Low MCA indicates the better approximation of the uncertainty in the data by the model. To better visualize this trend, we report a comparison of the R^2 values for these models in Appendix B Figure 6. Among the models, the Conv1D architecture provides minimally better accuracy in all cases.

Table 2: Comparison of accuracy metrics (MAE, RMSE, R^2) before and after adding wind forecast in the model during training.

Look-Ahead (Hours)	Model	MAE ↓		RMSE ↓		R^2 ↑	
		Before	After	Before	After	Before	After
24	MLP	0.31	0.34	0.47	0.46	0.76	0.74
	Conv1D	**0.29**	**0.28**	**0.42**	**0.39**	**0.79**	**0.82**
	LSTM	0.31	0.35	0.45	0.51	0.75	0.68
48	MLP	0.43	0.39	0.60	0.51	0.55	0.67
	Conv1D	**0.39**	**0.33**	**0.56**	**0.44**	**0.62**	**0.76**
	LSTM	**0.39**	0.41	0.58	0.58	0.56	0.59
72	MLP	0.46	0.41	**0.65**	0.54	0.48	0.63
	Conv1D	**0.45**	**0.38**	0.66	**0.50**	**0.51**	**0.73**
	LSTM	**0.45**	0.45	0.66	0.62	0.46	0.52

However, significant deviations from the actual water level patterns occur at peak water levels (above two feet) for longer forecast periods. This behavior is evident from the time series and scatter plots for the 72-hour forecast using the Conv1D model, as shown in Figure 1a. It is obvious that with an increase in the look-ahead period, the models are unable to predict high water levels accurately. Therefore, modifications to the models are required to capture unusual water level rises at least three days in advance. Among all the variables, water level exhibits a strong correlation with wind in terms of its components in X and Y directions. When forecasting 72 hours ahead using only 24-hour

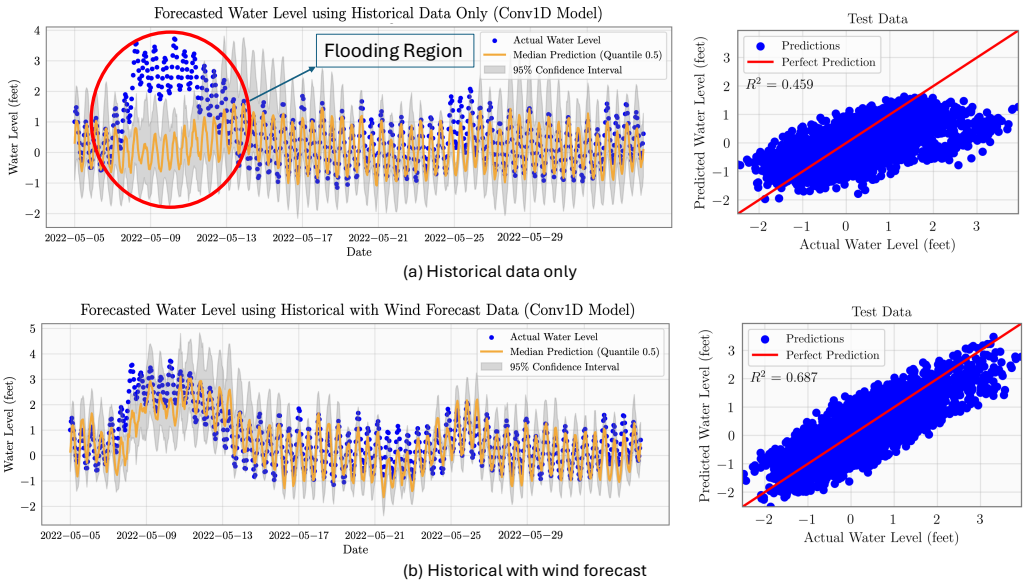

Figure 1: Time series plots for May 2022 and scatter plots for entire test data when (a) only historical data, and (b) historical with wind forecast are used for training the model.

historical data, the model does not have access to the wind for the target forecast period (72 hours later). If the wind remains within a normal range, the model performs well. However, if high winds occur at the forecast time, the model do not have any input information and remains completely unaware, leading to significant forecast inaccuracies. In order to overcome this limitation, we added forecasted wind decomposition as additional inputs to improve water level predictions for longer horizons. As a proof of concept, we utilize observed wind components data (derived from wind speed and its direction) from sensors as a proxy for forecasted wind to train and test the model. The revised model after adding wind forecast is now trained, and their predictive performances are summarized in Table 2 in terms of evaluation metrics. Interestingly, after incorporating the wind forecast, the Conv1D model achieves the highest accuracy, followed by MLP, which also shows significant improvement. However, the LSTM model does not exhibit notable performance gains due to the temporal misalignment between historical and forecasted data.

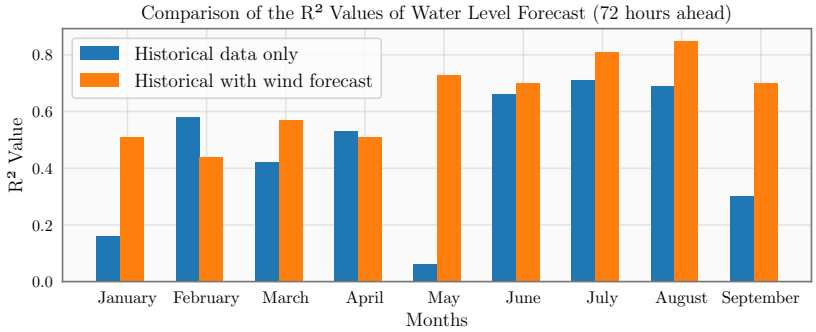

Figure 2: Monthly comparison of R^2 values for water level forecasts using historical data alone and historical data combined with wind forecasts.

Figure 1b presents the time series and scatter plots for 72 hours look-ahead using Conv1D model. Notably, the model now performs well in predicting high water levels (above 2 feet), a significant improvement compared to the model without the forecasted wind. Additionally, we compare the performance of the Conv1D model for monthly water level forecasts using only historical data versus historical data combined with wind forecasts. The analysis covers January to September 2022, as data for October to December are unavailable. Figure 2 indicates that for all months except February, forecast accuracy improves significantly with the inclusion of wind forecasts. For May 2022 when

flooding occurred in this area, the R^2 value increased dramatically from 0.06 to 0.73—a twelvefold improvement. We also examined the effect of adding precipitation forecasts in addition to wind forecasts; however, this did not lead to a significant improvement in model performance.

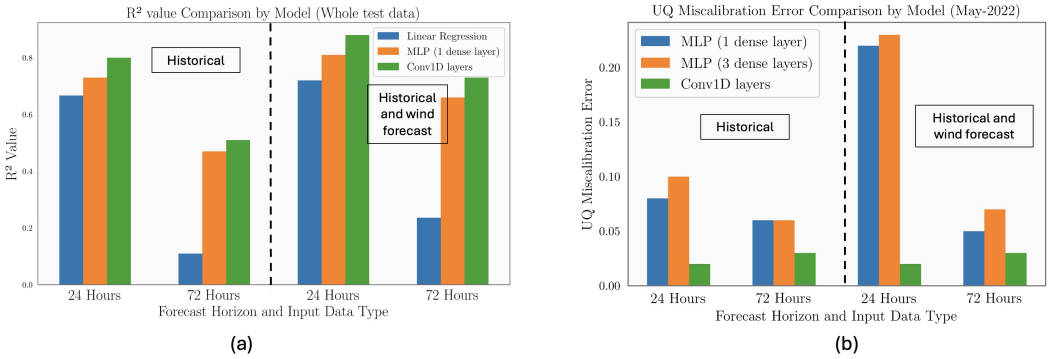

Figure 3: Comparison of the linear and non-linear model for 24 and 72 hours forcast using historical data only and historical data combined with wind forecast. (a) R^2 value for simple linear regression, MLP with one dense layer and Conv1D model (b) Unecratinty miscalibration error of MLP with one and three dense layers and Conv1D model.

To justify the necessity of a non-linear DL model, we compared the forecasting performance of several models across 24-hour and 72-hour prediction horizons, using two types of input: (i) historical water levels, and (ii) historical water levels combined with wind forecasts. Specifically, we evaluated a simple linear regression model, multilayer perceptron (MLP) with one and three dense layers using linear activation functions, and a Conv1D model that incorporates non-linear transformations. Predictive accuracy was assessed using the R^2 value, and the quality UQ was evaluated through miscalibration error. As summarized in Figure 3, the Conv1D model consistently achieved higher R^2 values (Figure 3a) and lower UQ miscalibration errors across (Figure 3b) both input configurations. These results demonstrate the advantage of non-linear modeling in capturing complex dependencies in the data and producing reliable uncertainty estimates.

5 Conclusion

Our model can forecast water levels at Little Neck Creek at Pinewood RD in Virginia Beach up to 72 hours (three days) in advance. This capability can play a critical role in managing and preventing flooding in the surrounding areas. The Public Works Operations team of the City of Virginia Beach provides flood stage classifications based on the water height at the gauge. Proactive forecasting products can help engineers, emergency managers, and decision-makers mitigate risks during adverse weather events. In addition, real-time monitoring of road conditions in the city can be enabled through computationally efficient predictive modeling of water levels at gauge locations. We plan to extend this work to seven-day-ahead forecasting using Transformer models. To support and improve decision making, we plan to develop a real-time dashboard to stream wind forecasts from the National Oceanic and Atmospheric Administration (NOAA) to display live forecast results. Moreover, Virginia Beach is a representative example of a coastal city where wind is often the main indicator of flooding. Our model's ability to forecast extreme events with uncertainty estimates is useful not just for this city, but also for other coastal areas around the world where wind plays a similar role in flooding.

Acknowledgments

Funding for this effort was provided by The Hampton Roads Biomedical Research Consortium as part of the efforts associated with the Joint Institute for Advanced Computing on Environmental Studies between Old Dominion University and Jefferson Laboratory. This manuscript has been authored by Jefferson Science Associates (JSA), operating the Thomas Jefferson National Accelerator Facility for the U.S. Department of Energy under Contract No. DE-AC05-06OR23177.

REFERENCES

Kasuni E Adikari, Sangam Shrestha, Dhanika T Ratnayake, Aakanchya Budhathoki, S Mohanasundaram, and Matthew N Dailey. Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions. *Environmental Modelling & Software*, 144:105136, 2021.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Thomas R Allen, Sridhar Katragadda, Yin-Hsuen Chen, Brian Terry, Joshua Baptist, Oguz Yetkin, Navid Tahvildari, Sunghoon Han, Blake Steiner, George McLeod, et al. A Digital Twin to Link Flood Models, Sensors, and Earth Observations for Coastal Resilience in Hampton Roads, Virginia, USA. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1388–1391. IEEE, 2023.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

Hamid Darabi, Ali Torabi Haghighi, Omid Rahmati, Abolfazl Jalali Shahrood, Sajad Rouzbeh, Biswajeet Pradhan, and Dieu Tien Bui. A hybridized model based on neural network and swarm intelligence-grey wolf algorithm for spatial prediction of urban flood-inundation. *Journal of hydrology*, 603:126854, 2021.

Francis Yongwa Dtissibe, Ado Adamou Abba Ari, Hamadjam Abboubakar, Arouna Ndam Njoya, Alidou Mohamadou, and Ousmane Thiare. A comparative study of Machine Learning and Deep Learning methods for flood forecasting in the Far-North region, Cameroon. *Scientific African*, 23: e02053, 2024.

Zhice Fang, Yi Wang, Ling Peng, and Haoyuan Hong. Predicting flood susceptibility using LSTM neural networks. *Journal of Hydrology*, 594:125734, 2021.

Roger Koenker. Quantile regression. *Cambridge Univ Pr*, 2005.

Vijendra Kumar, Hazi Md Azamathulla, Kul Vaibhav Sharma, Darshan J Mehta, and Kiran Tota Maharaj. The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management. *Sustainability*, 15(13):10543, 2023.

Jon Derek Loftis, David Forrest, Sridhar Katragadda, Kyle Spencer, Tammie Organski, Cuong Nguyen, and Sokwoo Rhee. StormSense: A new integrated network of IoT water level sensors in the smart cities of Hampton Roads, VA. *Marine Technology Society Journal*, 52(2):56–67, 2018.

Jon Derek Loftis, Molly Mitchell, Daniel Schatt, David R Forrest, Harry V Wang, David Mayfield, and William A Stiles. Validating an operational flood forecast model using citizen science in Hampton Roads, VA, USA. *Journal of Marine Science and Engineering*, 7(8):242, 2019.

Diana McSpadden, Steven Goldenberg, Binata Roy, Malachi Schram, Jonathan L Goodall, and Heather Richter. A comparison of machine learning surrogate models of street-scale flooding in Norfolk, Virginia. *Machine Learning with Applications*, 15:100518, 2024.

Binata Roy, Jonathan L Goodall, Diana McSpadden, Steven Goldenberg, and Malachi Schram. Forecasting Multi-Step-Ahead Street-Scale nuisance flooding using Seq2Seq LSTM surrogate model for Real-Time applications in a Coastal-Urban city. *Journal of Hydrology*, pp. 132697, 2025.

Jin Teng, Anthony J Jakeman, Jai Vaze, Barry FW Croke, Dushmanta Dutta, and SJEM Kim. Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental modelling & software*, 90:201–216, 2017.

Zening Wu, Yihong Zhou, Huiliang Wang, and Zihao Jiang. Depth prediction of urban flood under different rainfall return periods based on deep learning and data warehouse. *Science of The Total Environment*, 716:137077, 2020.

Lin Zhang, Huapeng Qin, Junqi Mao, Xiaoyan Cao, and Guangtao Fu. High temporal resolution urban flood prediction using attention-based LSTM models. *Journal of Hydrology*, 620:129499, 2023.

A CORRELATION HEATMAP

Figure 4 shows the correlation heatmap of water level with the wind speed & wind direction, and wind components in X and Y directions.

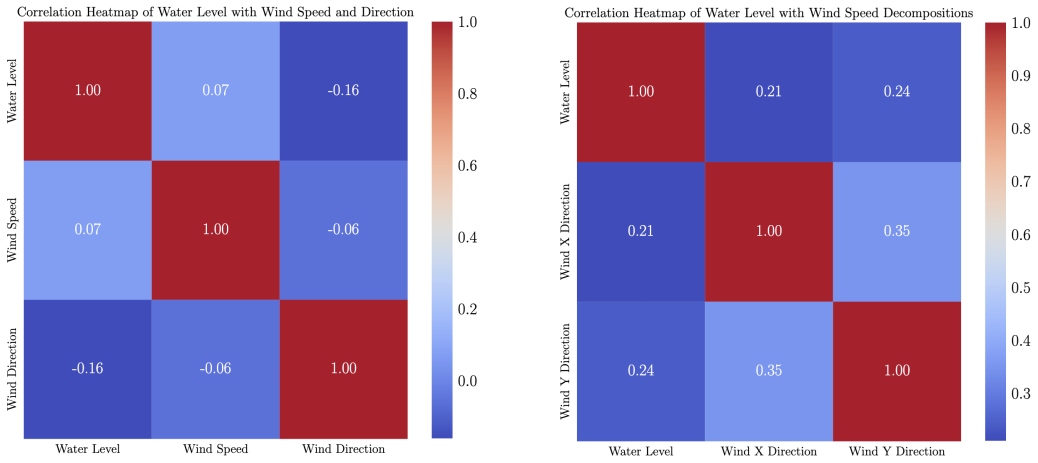

Figure 4: Correlations heatmap of the water level with wind speed and wind direction before decomposition (left) and water level with wind speed components in X and Y directions after decomposition (right).

B MODEL ARCHITECTURE AND ACCURACY COMPARISON

Table 3 summarizes the outcomes of the hyperparameter tuning, presenting the selected values for each model. Figure 5 illustrates the schematic of the DNN architectures employed in this study based on the selected hyperparameters. The MLP consists of three fully connected dense layers with linear activation function. The Conv1D model comprises two convolutional layers with max-pooling in the middle and a flattening layer before passing the data to a dense layer. Two LSTM layers are stacked for the LSTM model. Each model incorporates dropout layers with a rate of 0.1 to

Table 3: Summary of hyperparameter tuning results using Optuna framework.

Hyperparameter	**Range / Options Considered**	**MLP**	**Conv1D**	**LSTM**
Hidden Size / Filters	{32, 64, 128, 256}	128	128	128
Number of Layers	{1, 2, 3, 4, 5}	3	2	2
Kernel Size	Conv1D only: {3, 5, 7}	–	[3, 5]	–
Dropout Rate	{0.1, 0.2, 0.3, 0.4, 0.5}	0.1	0.1	0.1
Activation (Hidden Layers)	{ReLU, Tanh, SELU, Linear}	Linear	Tanh	Tanh
Activation (Output Layer)	{ReLU, Tanh, SeLU, Linear}	Linear	SeLU	ReLU
Learning Rate	Log-uniform: [1e-5, 1e-2]	0.0001	0.001	0.001
Batch Size	{64, 256, 512, 1024}	64	512	512

prevent overfitting. The output layer for each architecture is designed as a multi-output dense layer, where each neuron corresponds to a specific quantile used in DQR.

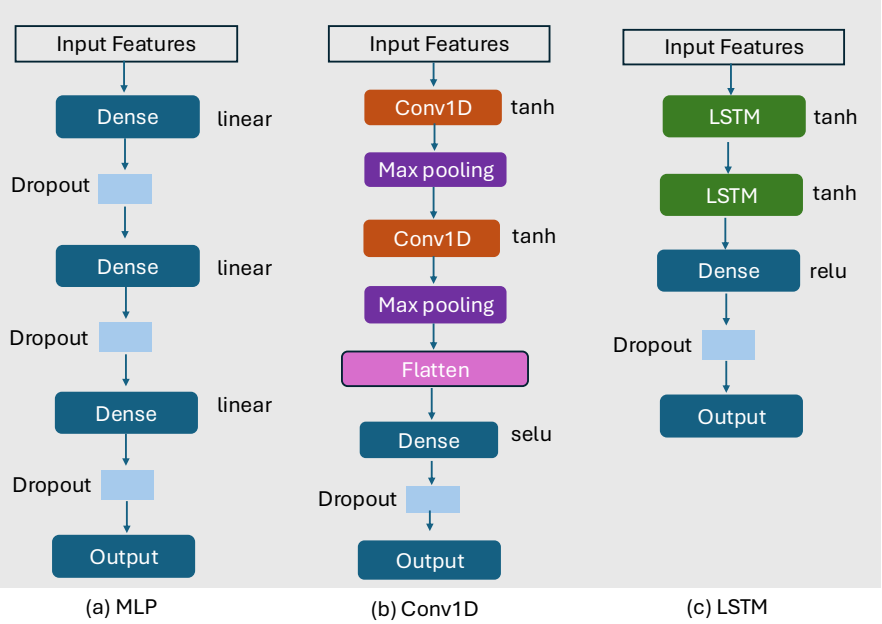

Figure 5: Schematic diagram of the Deep Neural Network (DNN) Architectures (a) Multi-Layer Perceptron (MLP), (b) 1D Convolutional Neural Network (Conv1D), and (c) Long Short-Term Memory (LSTM). Output layer of each architectures consists of predictions of the set of pre-defined quantiles.

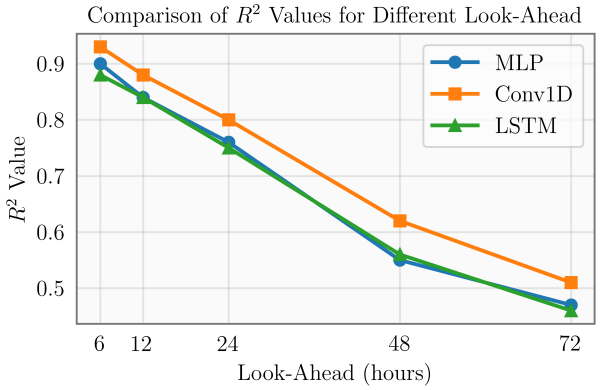

Figure 6: R^2 for all three models i.e., MLP, Conv1D, and LSTM with varying look-ahead using historical data only.

C Flood Control System of the City of Virginia

After each sensor installation, measurements are adjusted to the NAVD88 datum (North American Vertical Datum of 1988), a standardized reference system for measuring elevations in North America. Both National Weather Service (NOAA) and United States Geological Survey (USGS) gauges also provide data relative to this datum. Based on the water height recorded at the LNC gauge relative to NAVD88, the City of Virginia Beach defines four flood stages: Action, Minor, Moderate, and

Table 4: Flood stages and affected streets near the Little Neck Creek Gauge, VA Beach.

Flood Stage	Water Level (feet)	Affected Streets and Areas
Action	1.3’	Holly Rd begins flooding, South Bay Shore threatened.
Minor	2.0’	More than 6 inches of water on Holly Rd and South Bay Shore Dr. Pinewood Rd at Pump Station begins flooding. Oriole Dr and Kilbourne Ct threatened.
Moderate	2.5’	1 foot of water on Holly Rd and South Bay Shore. 6 inches on Oriole Dr and Kilbourne Ct. Bobolink, Bay Dr, and Pinewood begin flooding. Harris Teeter outfall system nears capacity.
Major	3.5’	More than 2 feet of water on Holly Rd and South Bay Shore. More than 1 foot on Oriole Dr, Pinewood Rd, Kilbourne Ct, and Bobolink. Arctic Ave begins flooding. Harris Teeter system at full capacity.

Major. Water height at this gauge is directly linked to flooding conditions in the surrounding streets, as summarized in Table 4.

Figure 7 illustrates how these flood stages impact different streets near the gauge. For example, the “Minor” flood stage may cause flooding on streets in the northwestern part of the area, while the “Action” flood stage may lead to flooding in the southern streets near the gauge. This system assists in making timely decisions regarding road closures and redirecting traffic patterns during flooding events. Additionally, Figure 7 shows a major flooding event in this area in May 2022, during which water levels reached 4 feet from May 5 to May 11.

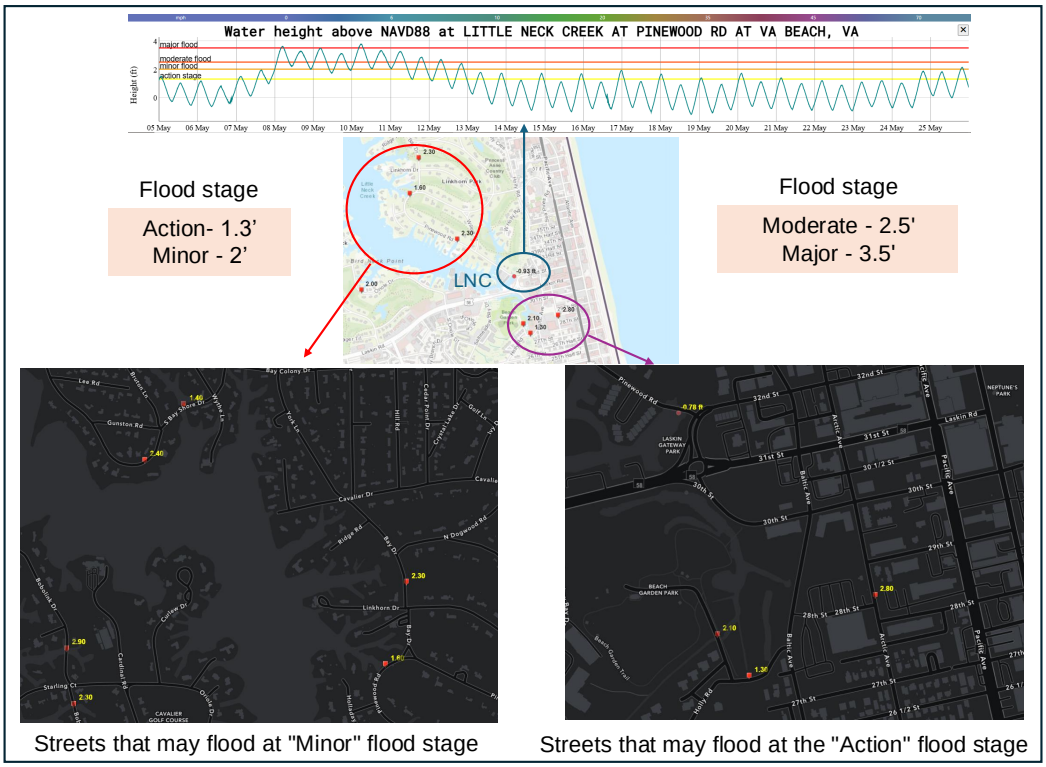

Figure 7: Schematic of the flood control system at the Little Neck Creek (LNC) gauge. The City of Virginia Beach defines four flood stages—Action, Minor, Moderate, and Major—which correspond to different flooding conditions in the surrounding streets. The top figure illustrates water level changes at LNC gauge during a significant flooding event in May 2022, when water levels reached up to 4 feet.