Uncertainty Quantified Machine Learning for Street Level Flooding Predictions in Norfolk, Virginia

Steven Goldenberg
TJNAF*
Newport News, VA 23606
sgolden@jlab.org

Diana McSpadden
TJNAF*
Newport News, VA 23606
dianam@jlab.org

Binata Roy
University of Virginia†
Charlottesville, VA 22904
br3xk@virginia.edu

Malachi Schram
TJNAF*
Newport News, VA 23606
schram@jlab.org

Jonathan L. Goodall
University of Virginia†
Charlottesville, VA 22904
goodall@virginia.edu

Heather Richter
Old Dominion University‡
Norfolk, VA 23529
hrichter@odu.edu

Abstract

Everyday citizens, emergency responders, and critical infrastructure can be dramatically affected by the flooding of streets and roads. Climate change exacerbates these floods through sea level rise and more frequent major storm events. Low-level flooding, such as nuisance flooding, continues to increase in frequency, especially in cities like Norfolk, Virginia, which can expect nearly 200 flooding events by 2050 [1]. Recently, machine learning (ML) models have been leveraged to produce real-time predictions based on local weather and geographic conditions. However, ML models are known to produce unusual results when presented with data that varies from their training set. For decision-makers to determine the trustworthiness of the model's predictions, ML models need to quantify their prediction uncertainty. This study applies Deep Quantile Regression to a previously published, Long Short-Term Memory-based model for hourly water depth predictions [2], and analyzes its out-of-distribution performance.

1 Introduction and Motivation

Coastal cities face combined forces of climate change: more frequent storms, increased precipitation, amplified storm surge and tidal action, sea level rise, and fluctuations in groundwater levels. These factors can lead to increased vulnerability and less resiliency to urban flooding [3]. Community members and city officials would benefit from a surrogate model capable of real-time prediction and fast simulation for what-if analysis, enabling them to understand potential disruptions in transportation, emergency management, and city services. Additionally, corresponding uncertainty estimations are essential for providing reliable decision support as they inform decision-makers when a data-driven model may be inadequately trained for climate or geographic conditions of interest.

The Hampton Roads region of Virginia has a total population of 1.7 million. The city of Norfolk is the second largest in the region, has the highest population density, and is home to over 245,000 people. Norfolk is a coastal city with 144 miles of waterfront, which enhances the quality of life and drives an economy that relies heavily on the Naval Station Norfolk and the Port of Virginia [4]. While access to water is an asset to the city, Norfolk is vulnerable to both pluvial and tidal flooding, which

*Thomas Jefferson National Accelerator Facility
†UVA Department of Civil and Environmental Engineering
‡ODU School of Community and Environmental Health

Tackling Climate Change with Machine Learning workshop at NeurIPS 2023

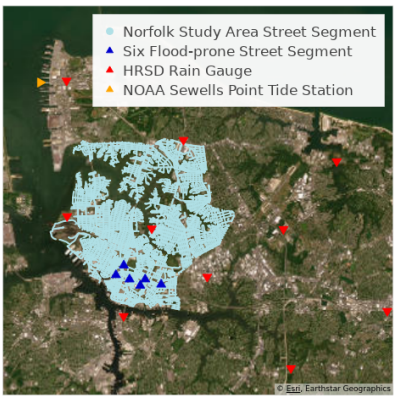

Figure 1: Plot of 16,923 Norfolk street segments in the full study area, and includes markers for the six flood-prone street segments, the Hampton Roads Sanitation District rain gauges, and the NOAA's Sewells Point tide station monitor.

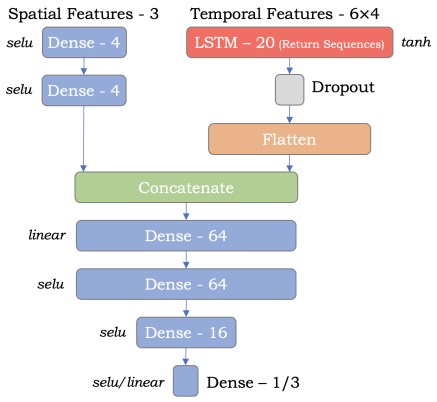

Figure 2: Model diagram for the Base and DQR models with activations in italics and output units following the dash. The final layer changes output size and activation for the Base/DQR models.

has increased in frequency and intensity due to climate change. As of 2018, the National Oceanic and Atmospheric Administration (NOAA) Sewells Point tide station documented the greatest change in relative sea level in the world of 1.45 feet in the last 100 years [3]. Additionally, since 1960, Norfolk has seen a 325% increase in nuisance flooding, lower levels of flooding not caused by extreme events or disasters [1]. With climate change, these trends are expected to continue, and the characteristics of tidal action, precipitation, soil saturation, and groundwater levels will continue to shift, leading to greater uncertainty in localized flooding predictions.

The importance of uncertainty quantification (UQ) for decision-relevant ML surrogate flooding models has been highlighted in the context of fluvial flooding prediction [5], storm-tracking [6], and flood risk [7]. ML flooding surrogate models have been developed for characterization of an entire rainfall event: forecasting flood risk [8, 9, 10], predicting occurrence of floods [11], estimating total accumulative water overflow [12], and predicting maximum water depth [13, 14, 15, 16]. However, these solutions do not include UQ. This work builds upon [17], which compared a physics-based 1-D/2-D high-resolution hydrodynamic model to a Random Forest (RF) surrogate model for street-scale flooding prediction in Norfolk and demonstrated RF inferences were approximately 3,000 times faster. Further research in [2] compared the performance of Recurrent Neural Network (RNN) surrogate models to [17]'s RF model and demonstrated equivalent error metrics. This work also employs RNNs as they support the inclusion of various UQ methods such as the Deep Quantile Regression (DQR) model presented here.

2 ML Methods and Datasets

2.1 Norfolk Rainfall Dataset

Data for this work includes both geospatial and weather components, which are standardized to zero mean and unit variance prior to model training. The geospatial data includes elevation (ELV), total wetness index (TWI) [18], and depth to water (DTW) measurements [19] for 16,923 street segments (or "nodes") in the Norfolk region shown in Fig. 1. 17 separate rainfall events ranging from June 5th, 2016 to August 20, 2018 make up the weather component of the dataset and include hourly rainfall, hourly tide level, 72 hour total rainfall, 2 hour total rainfall, and water depth estimations from a physics based Two-dimensional Unsteady FLOW (TUFLOW) model [20]. More information about the rainfall events can be found in Appendix A, Table 4, and the dataset is available for download on Hydroshare [21].

This dataset is split by events into training and testing sets that match previous work [17, 2]. In addition, we define three different node partitions; the first matches previous work on 6 flood-prone street segments (6FP), the second contains all nodes where all three geospatial features are within

Table 1: Accuracy results for the 6FP dataset. The mean and standard deviation in meters are computed from 16 random initializations of each model for the mean absolute error (MAE) and root-mean-square error (RMSE).

	Training		Testing	
Model	MAE	RMSE	MAE	RMSE
Base	0.0188 ± 0.0014	0.0463 ± 0.0024	0.0308 ± 0.0008	0.0625 ± 0.0015
DQR	0.0186 ± 0.0016	0.0437 ± 0.0032	0.0317 ± 0.0010	0.0649 ± 0.0024

the first and third quartiles (IQR), and the last contains all 16,923 nodes over the full study area (FSA). The IQR dataset was chosen to facilitate analysis of our model's accuracy and uncertainty quantification performance on out-of-distribution (OOD) data. More information on these data splits can be found in Appendix A, Table 5.

2.2 Base Long Short-Term Memory Model

Here we use the model architecture shown in Fig. 2, which was selected in [2] via a neural architecture search conducted on the six flood-prone street segments. A more detailed model description can be found in Appendix A, Table 3. As shown in Fig. 2, there are two input branches. A first input branch extracts features from the spatial inputs ELV, TWI, and DTW. The second captures information from the evolving tidal and pluvial events by employing a single Long Short-Term Memory (LSTM) layer using a four-time-step look back. Outputs from the branches are concatenated, and dense layers extract relevant features from the combined temporal and spatial information to produce a single time-step, i.e., one-hour look-ahead inference.

2.3 Deep Quantile Regression

Unlike a standard deep learning model, Deep Quantile Regression (DQR)[22] estimates conditional quantiles of the output. To do this, DQR uses the following alternative loss function:

$$\mathcal{L}(y_i, \hat{y}_i) = \max(\tau(y_i - \hat{y}_i), (\tau - 1)(y_i - \hat{y}_i)),$$

where τ is the desired quantile, and y_i and $\hat{y}_i$ are the label and prediction of the model respectively. In this paper, we define our desired quantiles as $\tau = [0.159, 0.5, 0.841]$, which provides a median prediction as well as quantiles that match expected proportions for one standard deviation assuming a $\mathcal{N}(0, 1)$ Gaussian distribution. We average the difference from the outer quantiles to the median to obtain a single standard deviation prediction. Computing uncertainty predictions using these quantiles allows for comparisons to other UQ methods and use of the Uncertainty Toolbox for calibration [23]. In order to make the quantile predictions, we updated the output layer of the base model to return three outputs with a linear activation function, instead of the single output with selu activation from the original model in [2].

3 Results

In order to verify our model performance matches the base model without UQ, and the results presented in [2], we trained 16 random initializations of the model with and without UQ. By training multiple initializations, we can verify the stability of our model architecture and calculate the mean and standard deviation for comparison metrics to determine whether there exist statistically significant differences between results. Table 1 includes the mean absolute error (MAE) and root-mean-square error (RMSE) for the training and testing event splits described in Appendix A, Table 4. While the two models have statistically significant differences (z-statistic > 2.5) for all metrics except the training MAE, it is clear that adding UQ through DQR only has a marginal effect on the average performance. Specifically, the largest difference is only 2.4 mm for the testing RMSE. This is expected as both model architectures are identical except for the output layer. Additionally, the DQR loss function for the median prediction is equivalent to the MAE loss function used by the base model.

Next, we trained 16 random initializations of the DQR model on the IQR dataset using the same 12 training events. We report accuracy and UQ calibration results calculated using the Uncertainty

Table 2: Results for the IQR dataset. The mean and standard deviation are computed from 16 random initializations of the DQR model for the mean absolute error (MAE), root-mean-square error (RMSE), root-mean-square calibration error (RMSCE) and miscalibration area (MA).

	Accuracy		UQ Calibration	
Dataset.	MAE	RMSE	RMSCE	MA
Training	0.0413 ± 0.0007	0.0633 ± 0.0012	0.0496 ± 0.0184	0.0454 ± 0.0172
Testing	0.0534 ± 0.0038	0.0828 ± 0.0089	0.0365 ± 0.0165	0.0314 ± 0.0158
FSA	0.3954 ± 0.2047	0.8744 ± 0.5291	0.1992 ± 0.0280	0.1766 ± 0.0248

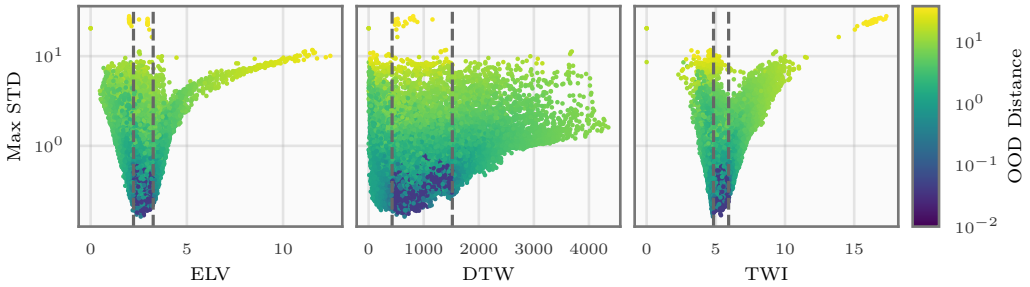

Figure 3: Scatter plots comparing the maximum returned standard deviation (uncertainty) over all timesteps for each street segment with the three geospatial features. The color of each point represents the minimum Euclidean distance between each street segment's geospatial features and the training set. Dotted vertical lines denote the first and third quartiles used to select the training data.

Toolbox python package [23] in Table 2. Similar to the 6FP results, the model performs well for in-distribution street segments on both the training and testing events with only minor signs of over-fitting. Additionally, the model is well-calibrated according to the root-mean-square calibration error (RMSCE) and miscalibration area (MA) metrics for the training and testing events.

Perhaps unsurprisingly, when we tested the DQR model on the FSA dataset over all events, we saw a significant reduction in accuracy and uncertainty calibration. However, the UQ calibration metrics degraded less than the accuracy and may still be useful. In fact, for the FSA dataset, the correlation between absolute prediction error (i.e. $|y - \hat{y}|$) and predicted standard deviations was 0.9545 ± 0.0442. Therefore, the standard deviations reported by the model are highly correlated with the true error, even when those errors are significantly larger than training errors.

Additionally, we examined the maximum uncertainty reported by the model for each street segment. To do this, we calculated the smallest distance between geospatial features for each street segment and the training set. These distances, which represent how OOD a street segment is from the training data, were used to color Fig. 3. We see increased standard deviations generally correlate with larger distances from the training geospatial features (lighter colors). Moreover, standard deviations increase more rapidly with TWI and ELV changes, which matches feature importance testing done in [17].

4 Conclusion

Given the expected sea level rise and the increasing storm vulnerability of Norfolk, Virginia, due to climate change, predicting the timing and severity of localized nuisance flooding will become increasingly relevant to the lives of residents. The DQR model presented in this paper allows for fast and accurate predictions of these flooding conditions while providing increased uncertainty for scenarios outside of the training distribution. These uncertainties are highly correlated with model error ($r = 0.95$) and, therefore, provide a reliable indicator for the trustworthiness of the model output. This is true even on novel OOD data, which suggests that the model may be transferable to other flood-prone areas like New Orleans, Louisiana, especially if trained on the full Norfolk study area.

References

[1] AG Burgos, BD Hamlington, Philip R Thompson, and Richard D Ray. Future nuisance flooding in Norfolk, VA, from astronomical tides and annual to decadal internal climate variability. *Geophysical Research Letters*, 45(22):12–432, 2018.

[2] 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. *arXiv preprint arXiv:2307.14185*, 2023.

[3] David Imburgia and Lucy Stoll. Resilient hampton. *Measuring and Evaluating Resilience*, 4, 2019.

[4] The City of Norfolk. Norfolk Resilience Strategy. `https://www.norfolk.gov/DocumentCenter/View/27257/Norfolk-Resilience-Strategy-?bidId=`, 2015. Accessed: 2023-09-12.

[5] Anouk Bomers and Suzanne JMH Hulscher. Neural networks for fast fluvial flood predictions: Too good to be true? *River Research and Applications*, 2023.

[6] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

[7] Anne Jones, Julian Kuehnert, Paolo Fraccaro, Ophélie Meuriot, Tatsuya Ishikawa, Blair Edwards, Nikola Stoyanov, Sekou L Remy, Kommy Weldemariam, and Solomon Assefa. AI for climate impacts: applications in flood risk. *npj Climate and Atmospheric Science*, 6(1):63, 2023.

[8] Marcel Motta, Miguel de Castro Neto, and Pedro Sarmento. A mixed approach for urban flood prediction using machine learning and gis. *International Journal of Disaster Risk Reduction*, 56:102154, 2021.

[9] 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.

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

[11] Xiaohan Li and Patrick Willems. A hybrid model for fast and probabilistic urban pluvial flood prediction. *Water Resources Research*, 56(6):e2019WR025128, 2020.

[12] Hyun Il Kim and Kun Yeun Han. Urban flood prediction using deep neural network with data augmentation. *Water*, 12(3):899, 2020.

[13] 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.

[14] Simon Berkhahn, Lothar Fuchs, and Insa Neuweiler. An ensemble neural network model for real-time prediction of urban floods. *Journal of Hydrology*, 575:743–754, 2019.

[15] Zifeng Guo, Joao P Leitao, Nuno E Simões, and Vahid Moosavi. Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *Journal of Flood Risk Management*, 14(1):e12684, 2021.

[16] Roland Löwe, Julian Böhm, David Getreuer Jensen, Jorge Leandro, and Søren Højmark Rasmussen. U-flood - topographic deep learning for predicting urban pluvial flood water depth. *Journal of Hydrology*, 603:126898, 2021.

[17] Faria T Zahura, Jonathan L Goodall, Jeffrey M Sadler, Yawen Shen, Mohamed M Morsy, and Madhur Behl. Training machine learning surrogate models from a high-fidelity physics-based model: Application for real-time street-scale flood prediction in an urban coastal community. *Water Resources Research*, 56(10):e2019WR027038, 2020.

[18] Keith J Beven and Michael J Kirkby. A physically based, variable contributing area model of basin hydrology/un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological sciences journal*, 24(1):43–69, 1979.

[19] Paul NC Murphy, Jae Ogilvie, Kevin Connor, and Paul A Arp. Mapping wetlands: a comparison of two different approaches for new brunswick, canada. *Wetlands*, 27(4):846–854, 2007.

[20] BMT WBM. Tuflow user manual. *BMT WBM*, 2016.

[21] Binata Roy, Steven Goldenberg, and Diana McSpadden. Input data for LSTM and seq2seq LSTM surrogate models for multi-step-ahead street-scale flood forecasting in Norfolk, VA. HydroShare. `http://www.hydroshare.org/resource/e5d6d32a320f4bcca679e0bf388c2bcc`, 2023.

[22] Roger Koenker. *Quantile regression*, volume 38. Cambridge university press, 2005.

[23] 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.

A Model and Dataset Information

Below, we present our model architecture where the base model matches [2]. The output shape and number of parameters are obtained with the Keras function *Model.summary()*. The layers for the temporal and spatial branches of the model are described separately with markers (*, †) to indicate their final output, which is passed to the mixture branch of the model. Values that have been changed for the DQR model are given in bold.

Table 3: Model architecture. Total params: 12,853/12,887

Model Branch	Layer	Output Shape	Activation	Param #	Connected to
Temporal	Input	(None, 4, 6)	—-	0	—-
	LSTM	(None, 4, 20)	tanh	2160	Input
	Dropout	(None, 4, 20)	—-	0	LSTM
	Flatten*	(None, 80)	—-	0	Dropout
Spatial	Input	(None, 3)	—-	0	—-
	Dense1	(None, 4)	selu	16	Input
	Dense2†	(None, 4)	selu	20	Dense1
Mixture	Concatenate	(None, 84)	—-	0	[*, †]
	Dense1	(None, 64)	linear	5440	Concatenate
	Dense2	(None, 64)	selu	4160	Dense1
	Dense3	(None, 16)	selu	1040	Dense2
	Dense Out	(None, 1/**3**)	selu/**linear**	17/**51**	Dense3

Table 4 is modified from [2] with event dates in chronological order. The Aug. 20, 2018 event is not used in this work due to its short duration which does not allow for adequate look-back for the temporal LSTM input layer.

Table 4: Rainfall Events

Event Dates	Duration (hrs)	Dataset
June 05-06, 2016	28	Train
July 30-31, 2016	34	Train
Aug. 09, 2016	16	Train
Sep. 02-03, 2016	28	Train
Sep. 19-21, 2016	60	Train
Oct. 08-09, 2016	37	Train
Jan. 01-02, 2017	23	Train
July 14-15, 2017	22	Train
Aug. 07-08, 2017	34	Train
Aug. 28-29, 2017	25	Train
Oct. 29-30, 2017	29	Test
May 06, 2018	24	Test
May 28-29, 2018	26	Test
June 21-23, 2018	37	Train
July 30, 2018	11	Train
Aug. 11, 2018	24	Test
~~Aug. 20, 2018~~	~~5~~	~~Train~~

Table 5: Data Split Information

Dataset	Samples	Events
Train 6FP	1,842	13
Test 6FP	522	4
Train IQR	648,691	13
Test IQR	183,831	4
FSA	6,667,482	17