AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR027038

**Key Points:**
- Surrogate machine learning models were trained for flood prediction using a high-resolution and high-fidelity physics-based model
- The surrogate model accurately emulated flooding depth and duration on streets simulated by the physics-based model
- A 3,000 times speedup was achieved with the surrogate model compared to the physics-based model, making it attractive for real-time decision support

**Correspondence to:**
J. L. Goodall,
goodall@virginia.edu

**Citation:**
Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., & Behl, M. (2020). 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*, e2019WR027038. https://doi.org/10.1029/2019WR027038

Received 31 DEC 2019
Accepted 6 SEP 2020
Accepted article online 12 SEP 2020



# 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

**Faria T. Zahura[1,2], Jonathan L. Goodall[1,2], Jeffrey M. Sadler[1,2,3], Yawen Shen[1,2], Mohamed M. Morsy[1,2,4,5], and Madhur Behl[2,6]**

[1]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA, [2]Link Lab, School of Engineering and Applied Science, University of Virginia, Charlottesville, VA, USA, [3]Now at United States Geological Survey, Middleton, WI, USA, [4]Irrigation and Hydraulics Engineering Department, Cairo University, Giza, Egypt, [5]Now at Dewberry, Fairfax, VA, USA, [6]Department of Computer Science, University of Virginia, Charlottesville, VA, USA

**Abstract** Mitigating the adverse impacts caused by increasing flood risks in urban coastal communities requires effective flood prediction for prompt action. Typically, physics-based 1-D pipe/2-D overland flow models are used to simulate urban pluvial flooding. Because these models require significant computational resources and have long run times, they are often unsuitable for real-time flood prediction at a street scale. This study explores the potential of a machine learning method, Random Forest (RF), to serve as a surrogate model for urban flood predictions. The surrogate model was trained to relate topographic and environmental features to hourly water depths simulated by a high-resolution 1-D/2-D physics-based model at 16,914 road segments in the coastal city of Norfolk, Virginia, USA. Two training scenarios for the RF model were explored: (i) training on only the most flood-prone street segments in the study area and (ii) training on all 16,914 street segments in the study area. The RF model yielded high predictive skill, especially for the scenario when the model was trained on only the most flood-prone streets. The results also showed that the surrogate model reduced the computational run time of the physics-based model by a factor of 3,000, making real-time decision support more feasible compared to using the full physics-based model. We concluded that machine learning surrogate models strategically trained on high-resolution and high-fidelity physics-based models have the potential to significantly advance the ability to support decision making in real-time flood management within urban communities.

## 1. Introduction

In recent years, flooding due to climate change and sea level rise has become a major concern of communities along the U.S. East coast (Ezer & Atkinson, 2014; Sweet & Park, 2014). The situation will be aggravated by projected increases in rainfall frequency and volume (Prein et al., 2017), as well as significant increases in relative sea level (Kulp & Strauss, 2019; Vermeer & Rahmstorf, 2009). Although most past studies assessing flood impacts primarily focused on rarely occurring, extreme storm events and their associated storm surge, the cumulative cost of relatively frequent, low-level flooding, also known as nuisance or recurrent flooding, can be greater than the extreme events (Moftakhari et al., 2017). Nuisance flooding caused by rain and tide adversely affects social and economic activities by disrupting transportation systems (Jacobs et al., 2018; Suarez et al., 2005) and compromising the performance of storm sewers (Flood & Cahoon, 2015). These events can be exacerbated by the joint occurrence of high tides and even moderate rainfall events, given the interplay between tide and rainfall in coastal communities (Lian et al., 2015; Shen et al., 2019; Sreetharan et al., 2017).

To assist decision makers in anticipating potential flooded areas and preemptively taking measures to mitigate socioeconomic disruptions caused by urban flooding, researchers and practitioners have focused on building accurate real-time flood prediction models. To simulate urban pluvial flooding, physics-based models are typically used. Applications of one-dimensional (1-D) hydraulic models such as SWMM (Rossman, 2004) and MIKE 11 (Danish Hydraulic Institute [DHI], 2017a) are common for simulating

flow through storm sewer networks and river channels. However, these models typically have less accurate lumped approaches for simulating overland flow (Mark et al., 2004). To more realistically simulate overland flow, two-dimensional (2-D) or 1-D/2-D dual drainage models like SOBEK (Deltares, 2018), MIKE FLOOD (Danish Hydraulic Institute [DHI], 2017b), HEC-RAS 2D, and Two-dimensional Unsteady Flow (TUFLOW) models (BMT WBM, 2016) can be used. However, 2-D models are computationally expensive (Leandro et al., 2009), difficult to calibrate, especially at a city scale (Caviedes-voullième et al., 2012), and typically require long running times (Lhomme et al., 2006), making them not yet suitable for real-time flood prediction (Bermúdez et al., 2018).

Reducing the complexity of these high-fidelity and computationally intensive physics-based models could provide a more efficient and practical solution for supporting real-time flood prediction. These reduced-complexity models are often referred to as surrogate models (Ong et al., 2003; Queipo et al., 2005; Razavi et al., 2012). For the purposes of this paper, surrogate models are grouped into two broad categories. The first category is lower-fidelity surrogates. Lower-fidelity surrogates are still physics-based but are less comprehensive compared to the original physics-based models and focus on the dominant physical processes relevant to a particular application (Alexandrov et al., 2001; Alexandrov & Lewis, 2001; Bermúdez et al., 2018; Madsen & Langthjem, 2001). The second category of surrogate models is response-surface models. A response-surface model uses machine learning (sometimes referred to as data-driven) models to approximate the original model response without directly simulating physical processes (Box & Wilson, 1951; Chen et al., 2020; James et al., 2018; Razavi et al., 2012; Simpson et al., 2001; Yan & Minsker, 2006). The latter approach of using machine learning models was found by Bermúdez et al. (2018) to be more precise in reproducing flood dynamics in a highly urbanized flat terrain and capable of gaining higher computational speedup factors compared to a low-fidelity surrogate model. Therefore, a machine learning surrogate approach was used in this study.

Machine learning or data-driven approaches have been used in water resources applications for over a decade (Khu et al., 2004). They have been used to approximate hydrodynamic and hydrologic models of river systems (Solomatine & Torres, 1996; Wolfs et al., 2015), flow in sewer systems (Wolfs & Willems, 2017), rating curves (Wolfs & Willems, 2014), and to calibrate rainfall-runoff processes (Khu et al., 2004). In flooding applications, machine learning techniques, such as $k$ means clustering and neural networks, were used by Chang et al. (2010) to develop a regional flood inundation system based on a 2-D noninertial overland flow model. A real-time inundation forecasting model was developed by Jhong et al. (2017) using a support vector machine (SVM) to approximate inundation depth simulated by a FLO-2D model at reference points. These depths were then spatially expanded using a geographic information system (GIS) to generate inundation maps during typhoons. Liu and Pender (2015) replicated the evolution of water depth and velocity from an ISIS2D model using flood hydrographs as input. Bermúdez et al. (2019) used discharge and tide levels in three streams as input data to approximate water depth and velocity at 25,000 control points simulated by the 2-D shallow water model Iber. The approximated water depth was interpolated to generate a flood inundation map. Studies done by Bermúdez et al. (2018) and Berkhahn et al. (2019) considered urban settings for real-time pluvial flood modeling. Bermúdez et al. (2018) emulated flood volume in four small regions, which was converted to flood depth using GIS. Berkhahn et al. (2019) approximated maximum water depth on streets using an artificial neural network (ANN) during 2 and 1 hr synthetic storm events simulated by the 1-D/2-D dual drainage model HE 2-D.

Despite the increased use of machine learning techniques in water resources and flood inundation modeling, in particular, there are still unresolved research questions. Some of the past studies considered flood hydrographs or discharge in a channel (river or drainage canal) to model flood inundation (Bermúdez et al., 2019; Chang et al., 2010; Liu & Pender, 2015). These studies mainly focused on the inundation that would occur if the capacity of the channel was exceeded due to a storm event and did not explain the flooding occurring inland due to capacity exceedance of storm water systems. The use of GIS algorithms was required for studies by Bermúdez et al. (2018) and Jhong et al. (2017) to estimate inundation depth following the use of a machine learning method. The street-level flood model developed by Berkhahn et al. (2019) in an urban setting solely employed machine learning to estimate water depth, but only the maximum water depth during a storm event was simulated. Time series characteristics of flooding to estimate not only flood depth but also duration were not addressed. Studies into real-time street-level flood modeling solely using machine learning and depicting the evolution of water depth on streets during storm events in an urban-

coastal environment, rather than only the flood peak, are lacking. Additionally, previous studies used ANNs (Berkhahn et al., 2019; Bermúdez et al., 2018; Chang et al., 2010) or SVMs (Bermúdez et al., 2019; Jhong et al., 2017; Liu & Pender, 2015) to develop surrogate models for flood prediction. Less work has focused on the use of Random Forest (RF) as surrogate models for street-scale flood prediction, while RFs have shown significant skill for other application areas in water resources for flood hazard risk assessment (Wang et al., 2015), analyzing topographic control on overland flow (Loos & Elsenbeer, 2011), forecasting reservoir inflow (Yang et al., 2017), and discharge (Yang et al., 2016), estimating flood severity (Sadler et al., 2018) and estimating soil moisture for flood risks and crop viability (Breen et al., 2020).

In this study, these research gaps are addressed by exploring the use of an RF as a surrogate model to emulate the response from a complex 1-D pipe/2-D overland hydrodynamic model using a TUFLOW model built and validated for a large portion of the coastal city of Norfolk, Virginia, USA. This model expands on work described by Shen et al. (2019) that built a TUFLOW model for a smaller portion of Norfolk. The TUFLOW model was capable of simulating street-scale urban-coastal flooding. RF surrogate models were developed in this study to approximate TUFLOW-simulated floods occurring on the streets. The study area had 16,914 street segments, covering nearly 700 km of roadways in the region. Topographic (e.g., elevation, topographic wetness index, and depth to water) and environmental (e.g., hourly rainfall, cumulative rainfall in previous hours, and hourly tide) features were used to predict TUFLOW-simulated water depths for these street segments during each hour of a storm event. Sixteen different storm events were used to train the RF models, and four other storm events were used to test the models. Two different strategies for training the RF surrogate models were explored: (i) training on only the most flood-prone street segments and (ii) training on all 16,914 street segments. The ability of the surrogate models to reduce computational expense for real-time flood warning applications was quantified as well. The goal of this research was to advance the ability to perform street-scale, real-time flood warning for urban coastal communities needed to address the pressing problem of nuisance flooding. Beyond this application area, the goal of this research was also to advance understanding of the mutual benefit of machine learning and physics-based modeling approaches that, when used in combination, offer a powerful means to simulate complex hydrologic systems.

## 2. Materials and Methods

### 2.1. Study Area

Norfolk, Virginia, USA (Figure 1) is located in the Hampton Roads region of Virginia and is the second most populous city in Virginia with significant commercial, military, and historical importance. Norfolk is the home to the world's largest naval bases, one of the two North Atlantic Treaty Organization (NATO) Supreme Allied Commander Transformation headquarters, and the second busiest port on the East Coast of the United States. Norfolk and the surrounding Hampton Roads region have been experiencing nuisance flooding due to low elevations, sea level rise, and regional land subsidence (Eggleston & Pope, 2013; Kleinosky et al., 2007). It is the second most vulnerable region in the United States to coastal flooding after New Orleans (Fears, 2012). The city is also actively pursuing innovative measures for improving its resilience to flooding. Norfolk is one of the Rockefeller 100 Resilient Cities (100 Resilient Cities, 2019) and has its own Office of Resilience. Considering its vulnerability to nuisance flooding, the vital role Norfolk plays in the national economy and security, and their active work in resilience, Norfolk is an ideal study area for this research.

### 2.2. Data

#### 2.2.1. Environmental and Topographic Data

The environmental data used to train the surrogate model comprise rainfall and tide level observations. Daily and 15 min rainfall observations were obtained from U.S. National Oceanic and Atmospheric Administration (NOAA) station at Norfolk International Airport (NOAA, 2018a) and Hampton Roads Sanitation District (HRSD) observation sites, respectively. Hourly tide levels referenced to the North American Vertical Datum (NAVD88) were obtained from NOAA's Sewells Point station (NOAA, 2018b). The data were obtained for the simulation duration 1 January 2016 to 31 December 2018. Locations of the rain gauges and tide gauge used in this study are shown in Figure 1. A 1 m Digital Elevation Model

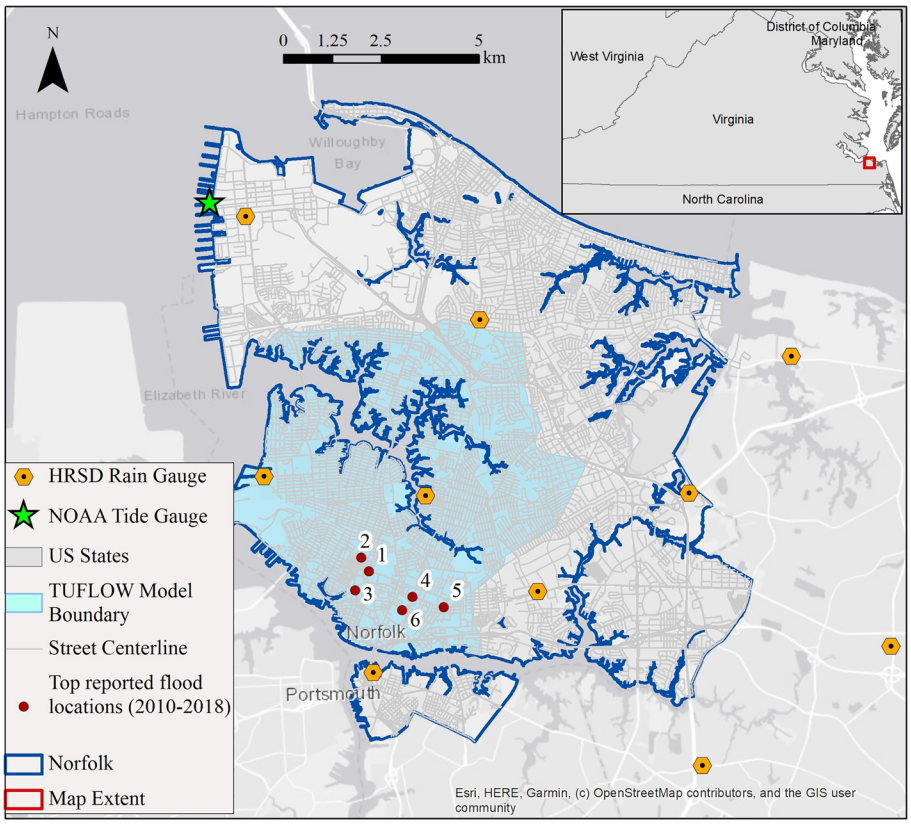


**Figure 1.** Map of the study area in Norfolk, VA, USA, showing rain and tide monitoring stations, the 1-D/2-D physics-based TUFLOW model boundary, and top flood-prone streets within the model domain.

(DEM) was obtained from the U.S. Geological Survey (USGS) through their National Elevation Dataset (NED) program (U.S. Geological Survey, 2016).

### 2.2.2. Physics-Based Model: TUFLOW

Street-level surface water depths were simulated using the physics-based TUFLOW model. TUFLOW solves 2-D equations for shallow water and free surface flow to simulate overland flow, and it is coupled with the 1-D hydrodynamic network software ESTRY (Syme, 2001) to simulate pipe flow. The 1-D pipe/2-D overland hydrodynamic flood model used in this study was an expanded version of the model described by Shen et al. (2019), and readers are referred to this reference for details on model construction, calibration, and evaluation process. The model used in this analysis covered an area of 56.4 $km^2$ in Norfolk, VA, as shown in Figure 1. The TUFLOW model used bare-earth DEM, which was modified using building footprints and heights. The study area had about 29,000 residential and commercial buildings. Implementing a large number of buildings can potentially increase model computing time and instability. Therefore, only buildings with areas greater than 500 $m^2$ were used to represent large commercial buildings and to reduce model complexity (Shen, 2020). The model simulated surface flooding at a 5 m spatial resolution and 1 hr time steps, which was then used to estimate water depths on street segments.

### 2.2.3. Roadway Network and Crowdsourced Data

A geospatial data set of the Norfolk roadway network was obtained from the city of Norfolk (City of Norfolk GIS Bureau, 2018a). To identify streets with high flood risk, flood reports cataloged in the city's System to Track, Organize, Record, and Map (STORM) from September 2010 to October 2018 were used. STORM data provided latitude and longitude of the locations where flood reports were made daily by city workers. The top six locations with the highest number of flood reports are shown in Figure 1. Flood reports were also obtained from the crowdsourced navigation app Waze (owned by Google), for September 2017 to August

**Table 1**
*Input Features for the RF Model*

| Input features | Feature abbreviation | Unit | Variability |
|---|---|---|---|
| | Environmental features | | |
| Total hourly rainfall | RH | mm | Spatial and temporal |
| Maximum 15 min rainfall in an hour | MAX15 | mm | Spatial and temporal |
| Cumulative rainfall in previous 2 hr | HR_2 | mm | Spatial and temporal |
| Cumulative rainfall in previous 72 hr | HR_72 | mm | Spatial and temporal |
| Hourly tide level from NAVD | TD_HR | m | Temporal |
| | Topographic features | | |
| Elevation | ELV | m | Spatial |
| Topographic wetness index | TWI | — | Spatial |
| Depth to water index | DTW | cm | Spatial |

2018 to validate flood occurrences. Waze allows users to report current conditions on streets, including flooded roads. It provides the exact time and location a report was made by a user to the nearest minute.

### 2.3. Machine Learning Model: RF Algorithm

RF is an ensemble machine learning algorithm for performing classification or regression that was first introduced by Breiman (2001). It makes predictions using a large collection of decorrelated decision trees with each decision tree in the RF algorithm learning from a random subset and input features. Predictions are made by combining responses from individual trees: mode of predicted classes for classification and average predictions for regression problems (Friedman et al., 2001). Decision trees are prone to overfitting to training data sets (Murphy, 2012). This issue is addressed in RF by introducing randomness in the training process and using enough trees to reduce overfitting. RF models can be optimized by adjusting hyperparameters such as the number of trees, number of features considered at each split, maximum depth of each decision tree, and minimum number of samples in a node before splitting. Additionally, another important quality of RF is its ability to estimate the importance of input features in predictions. The feature importance can be helpful to disregard unnecessary features and inform their significance in RF predictions.

Because the output (label) in this study was a continuous variable, surface water depth, an RF regression model was used. The Scikit-learn "ensemble.RandomForestRegressor" module in Python (Pedregosa et al., 2011; Scikit-learn Developers, 2018a) was used for RF regression. The scripts and data used for the analysis are available on HydroShare (Zahura, 2019a, 2019b, 2019c).

### 2.4. Input Data Preprocessing

Environmental and topographic data were used to generate the input features for the RF model described in Table 1. Rainfall data from HRSD stations at 15 min intervals were aggregated to generate four types of hourly rainfall features considering different types of rainfall events: hourly rainfall, maximum 15 min rainfall, cumulative rainfall during the previous 2 hr, and cumulative rainfall during the previous 72 hr. Maximum 15 min rainfall was important for those flood events when a large amount of rainfall occurred within a short period. Rainfall data during the previous 2 and 72 hr were used to account for antecedent moisture conditions and capacity exceedance of storm water systems due to rainfall occurrence immediately before the hour of interest and during the past few days, respectively. Because rainfall amounts vary from station to station, inverse distance weighted interpolation was performed for these rainfall features to estimate spatial variability within the study region. In an urban coastal environment, pluvial flooding can be exacerbated by the concurrent occurrence of high tide with rainfall. Therefore, hourly tide levels obtained from NOAA were also used as environmental features.

Three topographic features were used as model inputs: elevation, topographic wetness index (TWI), and depth to water (DTW). TWI and DTW were derived using the 1 m DEM. TWI, defined by Beven and Kirkby (1979) is


**Figure 2.** Map showing an example of road segments along the street centerlines for a small portion of the study area.

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right), \tag{1}$$

where α is the contributing area per unit contour length at a given point and tanβ is the local slope at that point in that catchment. TWI is a measure of the tendency of an area to accumulate runoff: High TWI values indicate a high potential for runoff accumulation.

DTW, proposed by Murphy et al. (2007), is an estimate of soil moisture conditions for each pixel (i.e., the smallest unit) of a DEM defined as

$$\mathrm{DTW} = \left[\sum \frac{dz_i}{dx_i} a_i\right] x_c, \tag{2}$$

where $\frac{dz_i}{dx_i}$ is the slope of a pixel $i$ in the landscape along the least-cost path to the nearest surface water pixel, $a$ is either 1 or $2^{0.5}$ depending on whether the path crosses the pixel parallel or diagonally to the pixel boundary, and $x_c$ is pixel size (Murphy et al., 2009). DTW approximates the elevation difference between a pixel in the landscape and the nearest surface water pixel along the least-slope path. Landscape pixels closer to surface water, in terms of both distance and elevation, have smaller values of DTW, suggesting both wetter soil and the path of least resistance for tidal flooding.

In terms of flooding impacts on roads, it was assumed that to restrict vehicular movement on the street due to flooding, it is sufficient to know the water depth at the deepest point along a given street segment. Therefore, road link centerlines were divided into 50 m roadway segments. Road width data were not available for all road links in the study domain. Because an average road lane width is 3.6 m in the United States (US Department of Transportation, 2014) and it was assumed that any street with missing width information has two lanes, the width for these road segments was assumed to be 7.2 m. A geospatial data set of road segments with 50 m lengths and 7.2 m widths was created using the ArcGIS software system (Esri, 2020), as shown in Figure 2. For different storm events at the location of each road segment, the mean

**Table 2**
*Training and Testing Events Used to Build the RF Models*

| Date | Daily rainfall (mm) | Maximum hourly rainfall averaged across stations (mm) | Train or test |
|---|---|---|---|
| 10/8/2016 | 188.98 | 29.738 | Train |
| 7/31/2016 | 177.29 | 34.29 | Train |
| 9/21/2016 | 99.82 | 21.84 | Train |
| 8/29/2017 | 99.82 | 13.97 | Train |
| 8/11/2018 | 94.49 | 27.94 | Test |
| 9/19/2016 | 77.22 | 23.37 | Train |
| 5/6/2018 | 65.28 | 18.03 | Test |
| 9/3/2016 | 61.21 | 13.21 | Train |
| 9/20/2016 | 60.45 | 10.67 | Train |
| 7/30/2018 | 59.94 | 18.54 | Train |
| 6/22/2018 | 57.91 | 15.49 | Train |
| 6/5/2016 | 53.59 | 28.19 | Train |
| 10/29/2017 | 53.09 | 7.06 | Test |
| 8/20/2018 | 52.32 | 35.31 | Train |
| 5/28/2018 | 47.75 | 13.46 | Test |
| 10/9/2016 | 45.72 | 19.30 | Train |
| 7/15/2017 | 45.47 | 28.96 | Train |
| 8/9/2016 | 44.70 | 15.49 | Train |
| 1/2/2017 | 43.94 | 8.38 | Train |
| 8/7/2017 | 43.94 | 16.76 | Train |

*Note*. Testing events are highlighted.

values of each input feature and the maximum value of TUFLOW-simulated hourly water depths were extracted. The underpasses in Norfolk have pump stations to prevent flooding. However, not enough information about these pump stations was available to represent them in the TUFLOW model. A limitation of the TUFLOW model was that it excluded pump information for road underpasses. As a result, these road underpasses and nearby roads were predicted to flood at levels deeper than might be expected in real life due to pumps at these underpasses. Therefore, these underpasses covering 3% of the road network analyzed in this study were excluded from the machine learning model. The end result was an RF model with 16,914 road segments capable of predicting hourly water depth on each road segment for a period before and after a storm event.

### 2.5. Model Training and Evaluation

From the daily rainfall data collected from Norfolk International Airport Station, the top 20 daily rainfall events were selected to build the flood prediction model (Table 2). The storm events were divided into an 80%/20% split for training and testing, respectively, which is a common approach in RF modeling (Agranoff et al., 2006; McFee & Lanckriet, 2010; Muchoney et al., 2000; Peng et al., 2004). The 16 training events included storm events of different durations and total rainfall depths. The events contained information from 2 or 3 hr before the storm event to several hours after the storm ended, resulting in a total of 375 hr of storm data in the training data set. For test event selection, the 20 daily observations were arranged in descending order and divided into four equal groups. Although we intended to choose one test event from each of the four groups, the group with the least rainfall amount did not have any Waze observations. Therefore, one test event was selected from each of the first two groups, and two events were chosen from the third group of daily rainfall observations. All of these test events included flood reports from Waze to verify flooded locations.

The workflow to prepare the input data set for the RF model is shown in Figure 3. This process was carried out for two different model training strategies. The first strategy was to train the surrogate model for just the six most flood-prone street segments in the study domain. These street segments were locations with the highest number of flood reports from the city of Norfolk and are shown in Figure 1. This more targeted strategy made an accurate model for these six locations but ignored network-wide flooding impacts. The sizes of the training feature and label data sets for this strategy were 2,250 rows and 8 columns, and 2,250 rows and 1

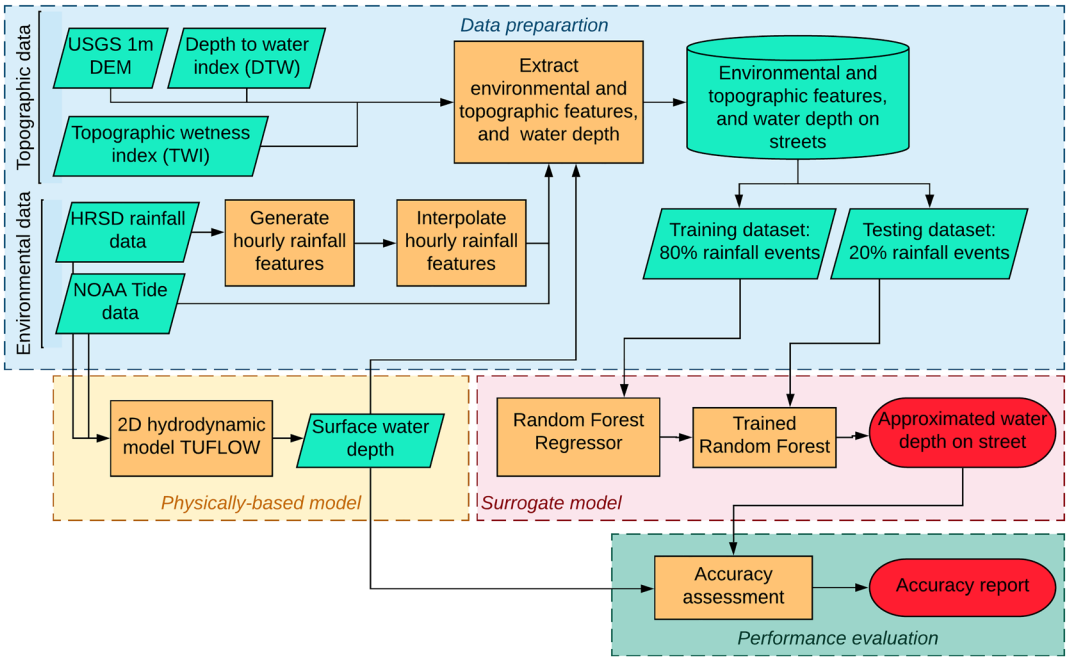


**Figure 3.** Workflow showing steps to generate the input data set for the RF models, approximate water depths using the models, and evaluate the models' performance.

column, respectively. The second strategy was to provide citywide flood predictions by training and testing an RF model for all 16,914 road segments within the TUFLOW model domain. The sizes of the training feature and label data sets for the second strategy were 6,342,750 rows and 8 columns, and 6,342,750 rows and 1 column, respectively. The RF model hyperparameters were optimized using GridSearchCV in Scikit-learn. GridSearchCV uses *k*-fold cross-validation to optimize model parameters while iterating over all possible combinations of provided parameter values. In *k*-fold cross-validation, the data set is partitioned into *k* number of groups. In each iteration, $(k-1)$ groups were used to train, and the remaining group was used to validate the model for a selected set of parameters. Each group was used once as the validation data set. The values of hyperparameters that maximized model accuracy were selected as the best parameter values. The main hyperparameters to be tuned in an RF model for improved performance were the number of trees (n_estimators) and the number of features to be considered at each split (max_features) (Scikit-learn Developers, 2018b). GridSearchCV was used on the training data set to optimize max_features for n_estimators varying between 1 and 300 with fourfold cross-validation. To determine the RF model sensitivity to different numbers of trees, the model was trained using n_estimators varying between 1 and 300 with the best performing max_features value from GridSearchCV. The other hyperparameters, such as the maximum depth of each tree and the minimum number of samples in a node before splitting, were set to default values "None" and two, respectively.

The training data set for the second strategy had samples with water depths ranges 0–0.1, 0.1–0.2, 0.2–0.3, and ≥0.3 m with ratios of 15.6:4.5:1.3:1; there were fewer samples in groups with high water depths, resulting in an imbalanced data set. As an imbalanced data set can interfere with model performance, different sampling techniques were tested to address this problem. The four sampling methods tested in this study were as follows: (i) Samples with water depths >0.2 m were assigned a "sample_weight" = 2, and the remaining samples had "sample_weight" = 1 while fitting the RF model (size of training feature and label data sets 6,342,750 rows and 8 columns, and 6,342,750 rows and 1 column, respectively); (ii) oversampling the minority group (i.e., water depths ≥0.3 m) to obtain a 1:1 ratio between sample groups with depths ≥0.3 and <0.3 m (no sample weight was assigned in this case; size of training feature and label data sets 12,227,103 rows and 8 columns, and 12,227,103 rows and 1 column, respectively); (iii) combination of methods (i) and (ii) (size of training feature and label data sets 12,227,103 rows and 8 columns, and 12,227,103 rows and 1 column, respectively); and (iv) the data set was used without any modification or sample

weight (size of training feature and label data sets 6,342,750 rows and 8 columns, and 6,342,750 rows and 1 column, respectively).

The performance of the surrogate model was assessed by comparing RF-predicted water depths on streets to TUFLOW-simulated water depths. Location-wise mean absolute errors (MAE) and root mean squared errors (RMSE) of RF-predicted and TUFLOW-simulated water depths for different test events were calculated to assess the performance of the RF model in predicting water depth at each road segment as

$$\mathrm{MAE_i(m)} = \frac{1}{n}\sum \left| y_{\mathrm{rf},\,h,\,i} - y_{\mathrm{tuflow},\,h,\,i} \right| \tag{3}$$

and

$$\mathrm{RMSE_i(m)} = \sqrt{\frac{\sum \left( y_{\mathrm{rf},\,h,\,i} - y_{\mathrm{tuflow},\,h,\,i} \right)^2}{n}}, \tag{4}$$

where $i$ indicates the location of water depth, $h$ represents an hour during the storm event, and $n$ is the total number of hours during the storm event, including 1 hr before and after the event. RF-predicted and TUFLOW-simulated water depths are represented by $y_{\mathrm{rf}}$ and $y_{\mathrm{tuflow}}$, respectively.

Pregnolato et al. (2017) found that, at a 0.30 m depth of water, a road becomes impassable for a passenger vehicle. Related work has shown that the limiting values of water depths on roadways are 0.10 and 0.30 m for high velocity and still water, respectively (Shand et al., 2011). From a decision-making perspective, road closure should occur at water depth above 0.30 m in most cases because the height of a car's air inlet ranges between 25 and 35 cm (AusRoads, 2008; Yin et al., 2016). Thus, three threshold values, 0.10, 0.20, and 0.30 m, were used to determine flood and nonflood locations within the road segment data set. These flood and nonflood designations were then used to assess the RF model's performance using the evaluation metrics precision, recall, and F1 score for the three threshold values. Precision and recall metrics were derived using true positive (TP), false positive (FP), and false negative (FN) values for each test event and flood threshold. TP, FP, and FN indicated the number of correctly predicted flood locations, TUFLOW-simulated nonflood locations incorrectly predicted as flood locations by the RF model, and TUFLOW-simulated flood locations incorrectly predicted as nonflood locations by the RF model, respectively. Recall, also known as sensitivity, was the proportion of TUFLOW-simulated flooding that was correctly predicted by the RF surrogate model and was defined as

$$\mathrm{Recall} = \frac{\text{Correctly predicted flood locations by RF}}{\text{Total simulated flood locations by TUFLOW}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

with values ranging from zero to one with one indicating that the model correctly predicted all flood locations. A low value of recall indicates underprediction of flooding, which may threaten human safety and cause property damage. In contrast, higher recall indicates most of the flooded areas were predicted correctly for necessary actions. Precision, also termed positive predictive value, was the proportion of total RF-predicted flooding that correctly matched with TUFLOW-simulated flooding and was defined as

$$\mathrm{Precision} = \frac{\text{Correctly predicted flood locations by RF}}{\text{Total predicted flood locations by RF}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6}$$

with values ranging from zero to one where one indicates that all predicted flood locations were correct. Low precision indicates overprediction of flooding, which may result in unnecessary actions in nonflood locations. F1 score is the weighted average of precision and recall and is defined as

$$\text{F1 score} = \frac{2*\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{7}$$

with values ranging from zero to one. With perfect precision and recall scores, the F1 score reaches a value of one. All performance metrics were calculated using the sklearn.metrics module (Scikit-learn Developers, 2018c) in Python.

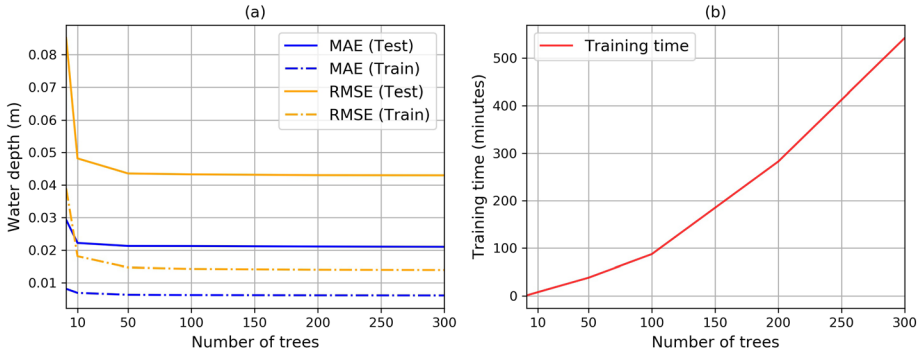


**Figure 4.** RF model sensitivity to a varying number of trees: (a) MAEs and RMSEs and (b) the training times with increasing number of trees.

For all model runs, two different machines were used: one for the TUFLOW simulations and another for the RF surrogate model training and testing. The TUFLOW model was run using a machine with 64GB RAM, four cores each running at 4.4 GHz, and two graphical processing units (GPU). The RF models, which run on a CPU rather than a GPU, were both trained and tested using a machine with 16GB RAM and a CPU with four cores, each running at 3.6 GHz. The run times for these models were recorded and compared to better understand the computational cost of each modeling approach.

## 3. Results and Discussion

### 3.1. Tuning RF Model Hyperparameters

Figures 4a and 4b show the improvement of MAEs and RMSEs for training and testing data sets and training time required with an increasing number of trees, respectively. Using GridSearchCV on max_features values ranging from 1 to 8, it was found that for varying n_estimators between 1 and 300, the best performing max_features value was 6 invariably. Therefore, max_features = 6 was used to assess RF model sensitivity to different numbers of trees (1, 10, 50, 100, 200, and 300). For the RF model with more than 50 trees, the improvement in MAEs and RMSEs was minimal while the required training time increased remarkably. Therefore, the n_estimators parameter in the RF model was set to 50 for further analysis.

### 3.2. Flood-Prone Streets Surrogate Model (RF Model 1)

Figure 5 compares water depth throughout the storm events modeled by both the RF surrogate and TUFLOW models for the six road segments used to develop the flood-prone streets surrogate model (RF Model 1). The MAEs and RMSEs of water depths at these six locations during the testing storm events are listed in Table 3. During the event on 11 August 2018, the surrogate model accurately matched the simulated time of peak water depth by the TUFLOW model. The peak water depth was overpredicted by RF Model 1 on five road segments by values ranging between 0.02 and 0.11 m. At Segment 2, the peak value was underpredicted by 0.011 m. As the storm passed, the surrogate model started to drain water out from the road surface, similar to the TUFLOW model. MAE and RMSE values for all locations were less than or equal to 0.042 and 0.061 m, respectively, showing a high predictive skill for the RF model. Predictions made for the 6 May 2018 event showed that at Segments 1 and 5, the surrogate model started to drain out the water before the physics-based model, resulting in higher MAE and RMSE values at those locations of 0.079 and 0.131 m, and 0.144 and 0.202 m, respectively. Although the MAEs and RMSEs are higher at Segments 1 and 5, the plots demonstrate that the surrogate model approximated the time of peak with good accuracy overall for the 6 May 2018 storm. Peak values during this event were slightly underpredicted on all segments, with an average difference of 0.072 m. During the 29 October 2017 event, less water accumulated on streets according to TUFLOW, which was reflected in the surrogate model output. The surrogate model overpredicted flood peak by an average of 0.068 m at Segments 1–4 and underpredicted flood peak by 0.036 m at Segments 5 and 6. The 28 May 2018 event had two distinct peaks. From the water depth versus time plot, it is evident that the RF surrogate was able to identify both the gradual rise in water level on the streets followed by the recession of water on all road segments. The highest peak during this event

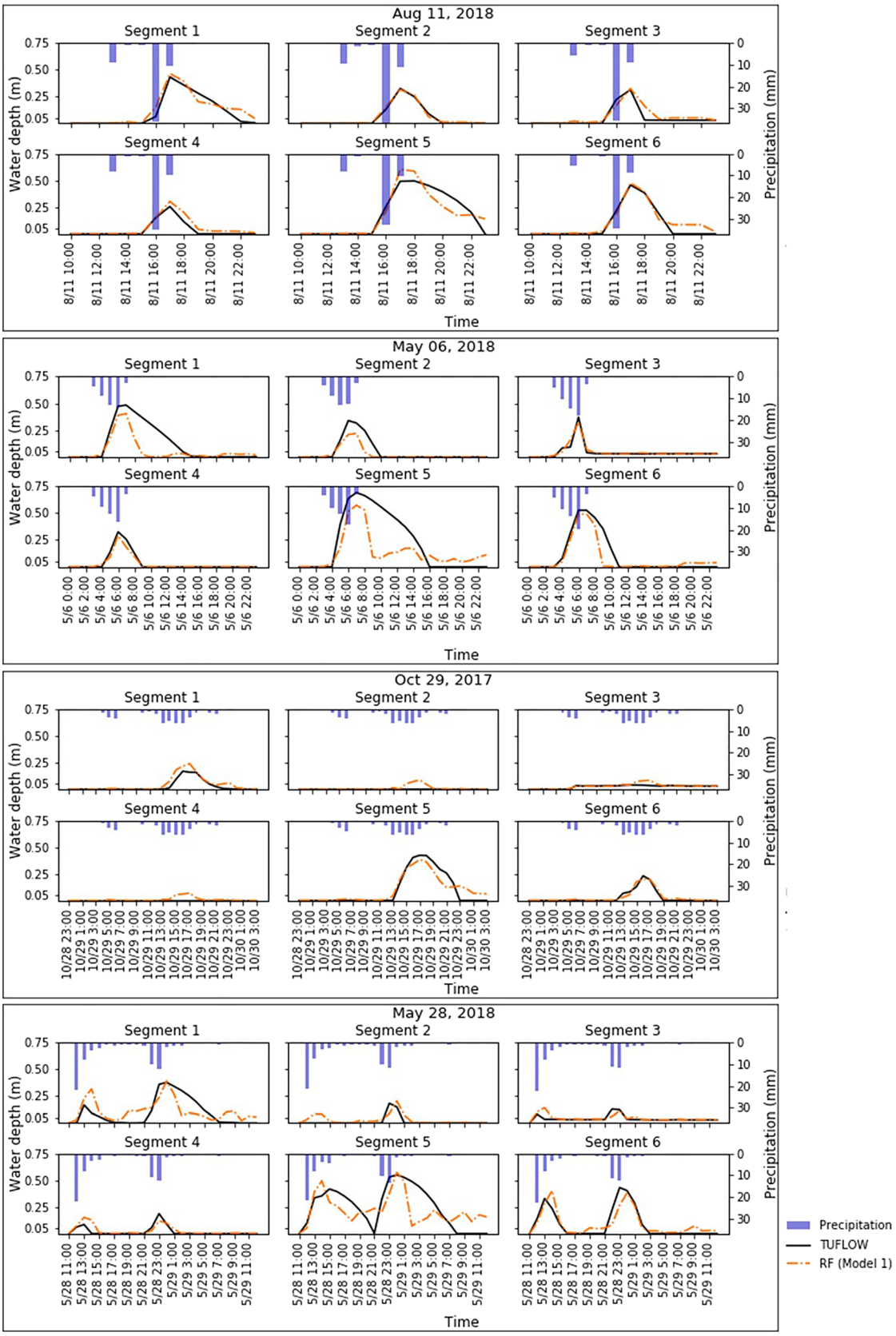


**Figure 5.** Water depth on the six road segments of RF Model 1 during the four test events.

**Table 3**
*MAEs (m) and RMSEs (m) of Water Depth for the Four Test Events and the Six Road Segments Used in RF Model 1*

| Events | Segment 1 | Segment 2 | Segment 3 | Segment 4 | Segment 5 | Segment 6 | Mean |
|---|---|---|---|---|---|---|---|
| | | | MAEs (m) | | | | |
| 29 Oct 2017 | 0.021 | 0.014 | 0.009 | 0.013 | 0.041 | 0.013 | 0.018 |
| 6 May 2018 | 0.079 | 0.026 | 0.010 | 0.009 | 0.144 | 0.041 | 0.051 |
| 28 May 2018 | 0.077 | 0.022 | 0.011 | 0.015 | 0.129 | 0.041 | 0.049 |
| 11 Aug 2018 | 0.031 | 0.010 | 0.033 | 0.023 | 0.042 | 0.019 | 0.026 |
| | | | RMSEs (m) | | | | |
| 29 Oct 2017 | 0.035 | 0.029 | 0.016 | 0.024 | 0.061 | 0.019 | 0.031 |
| 6 May 2018 | 0.131 | 0.060 | 0.020 | 0.021 | 0.202 | 0.086 | 0.087 |
| 28 May 2018 | 0.098 | 0.037 | 0.026 | 0.033 | 0.162 | 0.062 | 0.070 |
| 11 Aug 2018 | 0.046 | 0.014 | 0.052 | 0.038 | 0.061 | 0.026 | 0.039 |

was overpredicted by an average of 0.02 m at Segments 1, 2, and 5 and underpredicted by an average of 0.046 m at Segments 3, 4, and 6. Overall, these results suggested that RF Model 1 was able to learn from different topographic and environmental information provided to emulate the responses from the 2-D physics-based TUFLOW model. Although the surrogate model showed some deviations from the TUFLOW model for some segments and events, most often, it was able to estimate the time of peak and peak water depth successfully with low overall MAE and RMSE of 0.036 and 0.057 m, respectively. Precision and recall scores used in the second, citywide model (RF Model 2) were not calculated for RF Model 1 because these metrics are less relevant when comparing only six road segments.

### 3.3. Citywide Surrogate Model (RF Model 2)

#### 3.3.1. Effect of Sampling Techniques on RF Model Performance

The effect of different sampling techniques on RF Model 2 performance in terms of MAE, RMSE, precision, and recall is shown in Figure 6. The overall MAE and RMSE of the test data set were largest when the training data set was used without any modification, and assigned sample weight was "None." Assigning a sample weight of two to samples with water depth >0.2 m while fitting the model showed minor improvement. Maximum improvement in performance in terms of RMSE and MAE was obtained when the minority group with water depth ≥0.3 m was oversampled. Similarly, Figure 6b demonstrates that for groups with water depth >0.2 and >0.3 m, recall values increased by 32% and 57%, respectively, due to oversampling the minority class, while precision value improved by 7% and 3%, respectively. For the water depth >0.1 m group, precision improved by 10% with minimal improvement in recall value due to oversampling. Overall, oversampling the minority class in combination with assigned sample weight showed maximum improvement in model performance. Therefore, this technique was used for further analysis.

#### 3.3.2. Flood Depth Estimation

Figure 7 compares results from RF Model 2 to both the TUFLOW model and RF Model 1 for the six road segments with the highest number of flood reports. Flood reports from Waze at these locations are also

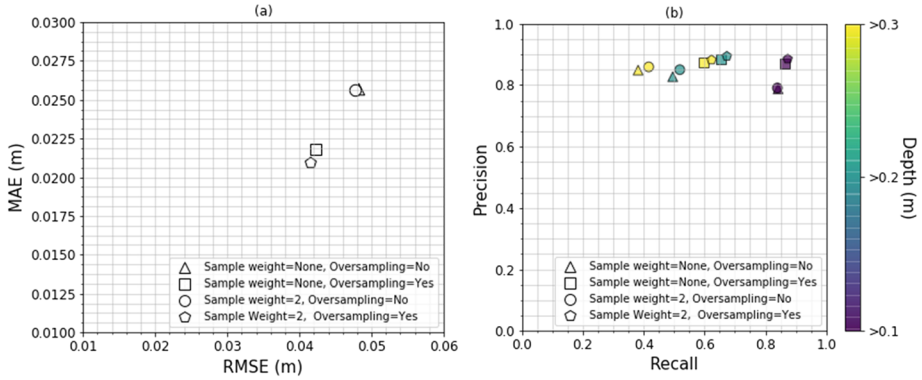


**Figure 6.** Effect of sampling techniques on RF model performance.

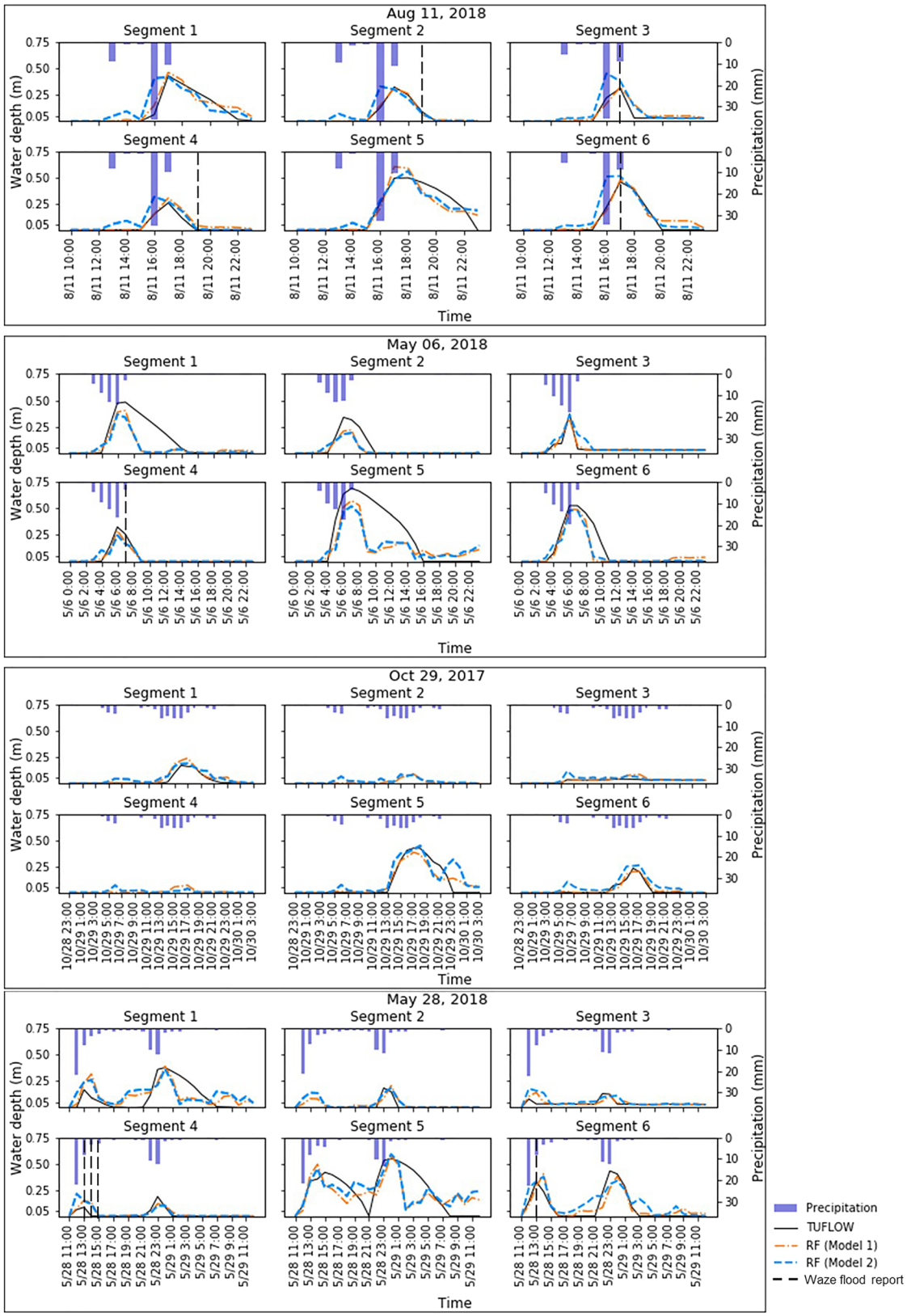


**Figure 7.** Comparison between RF Models 1 and 2 predictions and TUFLOW-simulated water depth on the six most flood-prone road segments during the four testing events. Vertical dashed lines represent occurrences of flood reports from Waze.

**Table 4**
*Water-Depth MAEs (m) and RMSEs (m) for the Six Road Segments From RF Model 2, Including Average and 90th Percentile Errors for 16,914 Segments and Four Test Events*

| Events | Segment 1 | Segment 2 | Segment 3 | Segment 4 | Segment 5 | Segment 6 | Average errors | 90th percentile errors |
|---|---|---|---|---|---|---|---|---|
| | | | | MAEs (m) | | | | |
| 29 Oct 2017 | 0.028 | 0.026 | 0.015 | 0.011 | 0.065 | 0.042 | 0.014 | 0.023 |
| 6 May 2018 | 0.083 | 0.031 | 0.016 | 0.016 | 0.153 | 0.042 | 0.014 | 0.025 |
| 28 May 2018 | 0.099 | 0.023 | 0.022 | 0.019 | 0.124 | 0.060 | 0.010 | 0.019 |
| 11 Aug 2018 | 0.073 | 0.048 | 0.066 | 0.053 | 0.039 | 0.066 | 0.030 | 0.053 |
| | | | | RMSEs (m) | | | | |
| 29 Oct 2017 | 0.034 | 0.037 | 0.024 | 0.019 | 0.105 | 0.053 | 0.026 | 0.035 |
| 6 May 2018 | 0.135 | 0.065 | 0.039 | 0.037 | 0.203 | 0.083 | 0.038 | 0.047 |
| 28 May 2018 | 0.118 | 0.046 | 0.042 | 0.043 | 0.164 | 0.083 | 0.029 | 0.038 |
| 11 Aug 2018 | 0.129 | 0.078 | 0.096 | 0.073 | 0.054 | 0.110 | 0.055 | 0.079 |

shown in the figure for context. Table 4 lists MAEs and RMSEs for RF Model 2 at the six flood-prone locations and the average and 90th percentile errors for 16,914 road segments. Comparing Table 3 to Table 4, MAEs and RMSEs were higher for RF Model 2 predictions, which was to be expected because RF Model 1 was specifically developed for those six locations whereas RF Model 2 learned from the topographic and environmental data for the 16,914 different road segments. MAE and RMSE differences were relatively small: no more than 0.047 and 0.084 m across the events and segments with an average of 0.015 m and 0.023 m, respectively. This demonstrated that, while targeted training improved accuracies for the most flood-prone streets, a single, citywide model trained on all 16,914 street segments also yielded accurate results in terms of the decision-making threshold values. In practice, an ensemble approach may be most appropriate that combines targeted RF Model 1 for the most flood-prone streets with RF Model 2 for citywide coverage to best inform decision makers.

While RF Model 2 produced overall results with low MAEs and RMSEs according to average values in Table 4, water level peak timing and depths notably differed from the targeted RF Model 1 values. During the 11 August 2018 event, RF Model 2 predicted the occurrence of the peak water depth 1 hr earlier than the TUFLOW model at Segments 2–4, potentially due to the significant rainfall at that hour. At Segments 1, 5, and 6, the time of peak predicted by RF Model 2 coincided with that of the TUFLOW simulation. Peak water depths were overpredicted by RF Model 2 at five segments with the difference between peak values from RF Model 2 and TUFLOW ranging from 0.006 to 0.14 m, whereas the peak value was underpredicted at Segment 1 by 0.01 m. TUFLOW-simulated water depths were significantly underpredicted by RF Model 2 at Segments 1 and 5 with a difference between peak values of 0.115 and 0.171 m, respectively, on 6 May 2018. In addition, RF Model 2 predicted to drain water out of these two segments before the TUFLOW-simulated time, resulting in high MAEs and RMSEs. Nevertheless, predicted maxima (Figure 7) was above 0.30 m, which is sufficient to make decisions about road closures, as discussed in section 2.5. During the 29 October 2017 event, less water accumulated on Segments 2–4, according to TUFLOW, which

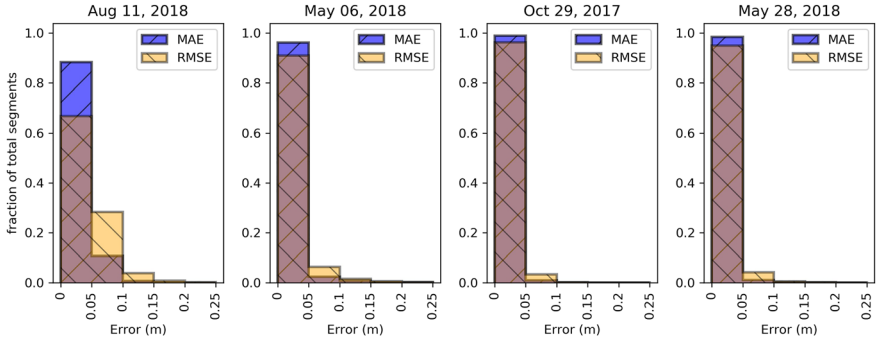


**Figure 8.** Histogram of MAE and RMSE between RF Model 2 and the physics-based TUFLOW model for water depth predictions at different road segments during test events.

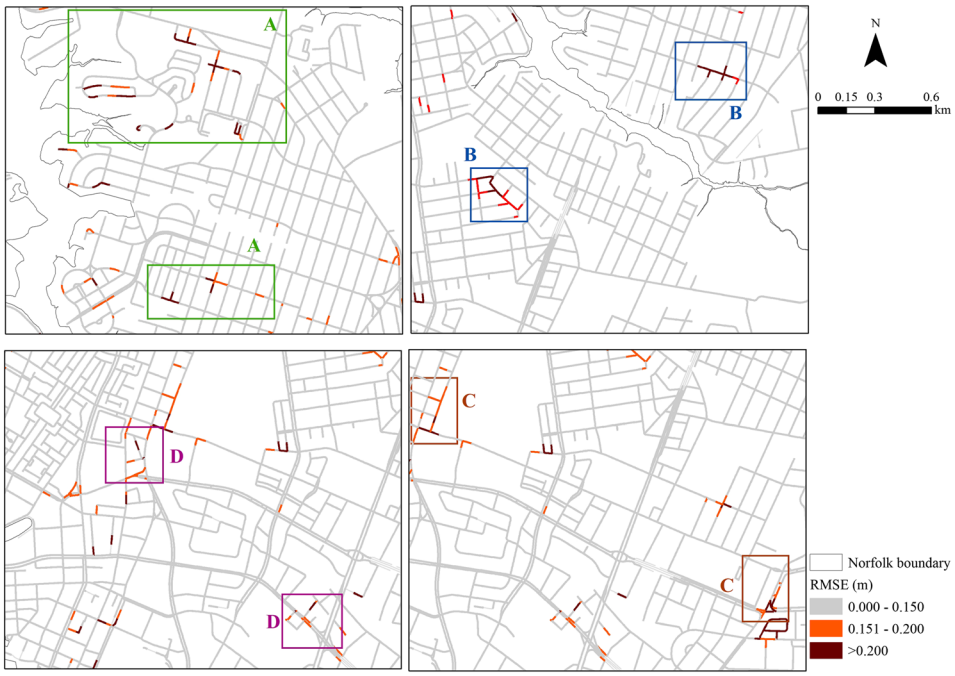


**Figure 9.** Map showing RMSE between RF Model 2 predicted and the physics-based model TUFLOW-simulated water depth for each road segment during the test events, where possible missing pipes in the TUFLOW model (A), possible incorrect parameterization of pipes (B), locations where RF Model 2 underpredicted but correctly predicted above 0.3 m threshold (C), and locations where RF Model 2 overpredicted flood levels (D).

was reflected in the surrogate model output. At Segments 1, 5, and 6, peak water depths were overpredicted by RF Model 2, with an average difference of 0.024 m. The two flood peaks simulated by the TUFLOW model on 28 May 2018 were also identified by the RF surrogate models; however, the first peak was overpredicted by RF Model 2 on all the street segments with an average peak difference of 0.09 m. The second peak during this event was underpredicted by RF Model 2 on five road segments with an average peak difference of 0.043 m, whereas an overprediction of 0.04 m occurred at Segment 5.

Crowdsourced flooding observations from the Waze app represent a binary "yes" or "no" value (not a water depth); however, they can aid in validating occurrences and durations of flooding in the road network. For example, the flood reports from Waze on 11 August 2018 (Segments 2, 3, and 6), 6 May 2018 (Segment 4), and 28 May 2018 (Segment 6) corresponded to periods of significant street flooding. However, the flood report on 11 August 2018 at 7:00 p.m. was made when water completely drained out of Segment 4 according to the TUFLOW simulation and RF prediction. Importantly, Waze reports should be considered when the predicted flood times and places from the RF surrogate models differed from the TUFLOW model. For example, three Waze reports were made at 1:00, 2:00, and 3:00 p.m. on 28 May 2018, close to Segment 4. However, the TUFLOW simulation suggested there was no water on that segment or nearby streets at 2:00 or 3:00 pm on 28 May 2018. On the other hand, the surrogate model predicted a water depth of 0.12 m at 2:00 p.m. for this event. Additionally, RF-predicted peak water depth at Segment 4 on 28 May 2018 was higher than the TUFLOW simulation. The RF model predicted the flooding observed by one or more Waze users, but the physics-based TUFLOW model did not.

The performance of RF Model 2 across the 16,914 road segments for the four test events was measured using MAE and RMSE computed at each road segment. Figure 8 shows histograms of predictive error in water depth for the four test events with average and 90th percentile MAEs and RMSEs listed in Table 4. The 11 August 2018 event had the highest MAE and RMSE, which is expected because it was the most extreme rainfall event in the test data set. Approximately 98% of the road segments had MAEs $< 0.084$ m and RMSEs $< 0.13$ m during this event. For the other three test events, >96% and 1–3% of the locations had MAEs $< 0.05$ m and within 0.05–0.10 m range, respectively. For all events, road segments with an MAE $> 0.10$ m was <1%. These results suggested that the surrogate model was effectively emulating the

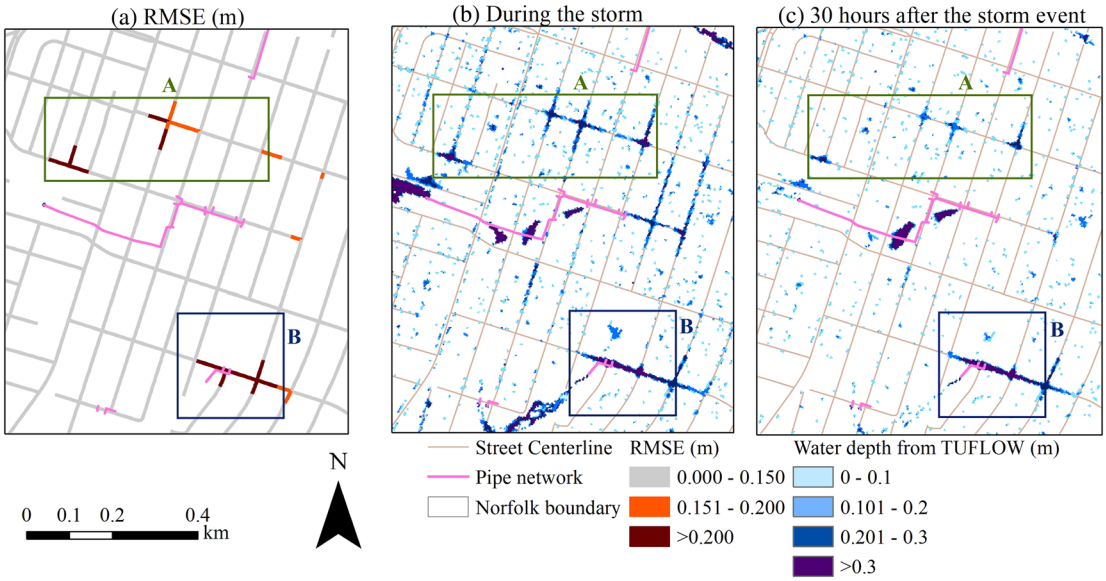


**Figure 10.** Locations with RMSE > 0.2 m and pipe network from the city of Norfolk. (A) indicates areas where there are possible missing pipes in the TUFLOW model while (B) indicates areas where there is possible incorrect parameterization of the pipes.

TUFLOW-simulated flood depth at most of the street segments in the study domain with low MAEs and RMSEs.

The use of the RF model as a surrogate might be used as a way to detect places where the TUFLOW model could be better parameterized or configured. Places where the emulation model consistently performed poorly might indicate locations where the physics-based model was behaving anomalously. Machine learning has long been used for anomaly detection (Lane & Brodley, 1997). Figures 9 and 10 highlight locations where RF Model 2 did not perform well. Although the number of locations with inaccurate predictions increased for larger events, clusters of locations existed where the RF surrogate consistently resulted in inaccurate results for all the events. Upon analyzing RF performance, 133 road segments had RMSEs > 0.2 m during any of the test events. At 71 of these segments, water drained much more slowly compared to other similar locations in the model domain. These road segments were analyzed in conjunction with storm sewer pipe data that were collected from the city of Norfolk (City of Norfolk GIS Bureau, 2018b) to build the TUFLOW model. Some locations, like the ones labeled "A" in Figures 9 and 10, had no 1-D storm water infrastructure in the TUFLOW model on streets where we would expect that infrastructure. When we examined street-view imagery from Google Maps, we found that there were indeed storm water inlets on the sides of the roads. The RF model, therefore, brought a potential problem in the physics-based model to our attention, and with some investigation, we were able to confirm that indeed the TUFLOW model was missing some information that is in the physical system.

Another group of locations that drained unusually slow (compared to the rest of the TUFLOW model) had 1-D storm water infrastructure represented at those locations (labeled "B" in Figures 9 and 10). Because the water drainage was slower at these locations than the rest of the model, and no indications of flooding from the WAZE data, we suspected that the 1-D storm water infrastructure might be incorrectly parameterized in the TUFLOW model at or near these locations. Investigating the pipe information at these locations, we found that the downstream invert levels of these pipes were higher than the upstream invert levels, which prohibited flow through these pipes in TUFLOW simulation. Future work conducting field surveys or deploying water level sensors to detect water depth during storm events at those locations would help to verify further and better understand what is occurring at these locations.

In addition to locations that were consistently and unusually slow to drain, there were another 50 locations that had high RMSE. At these locations, peak water depths during the storm events were simulated between

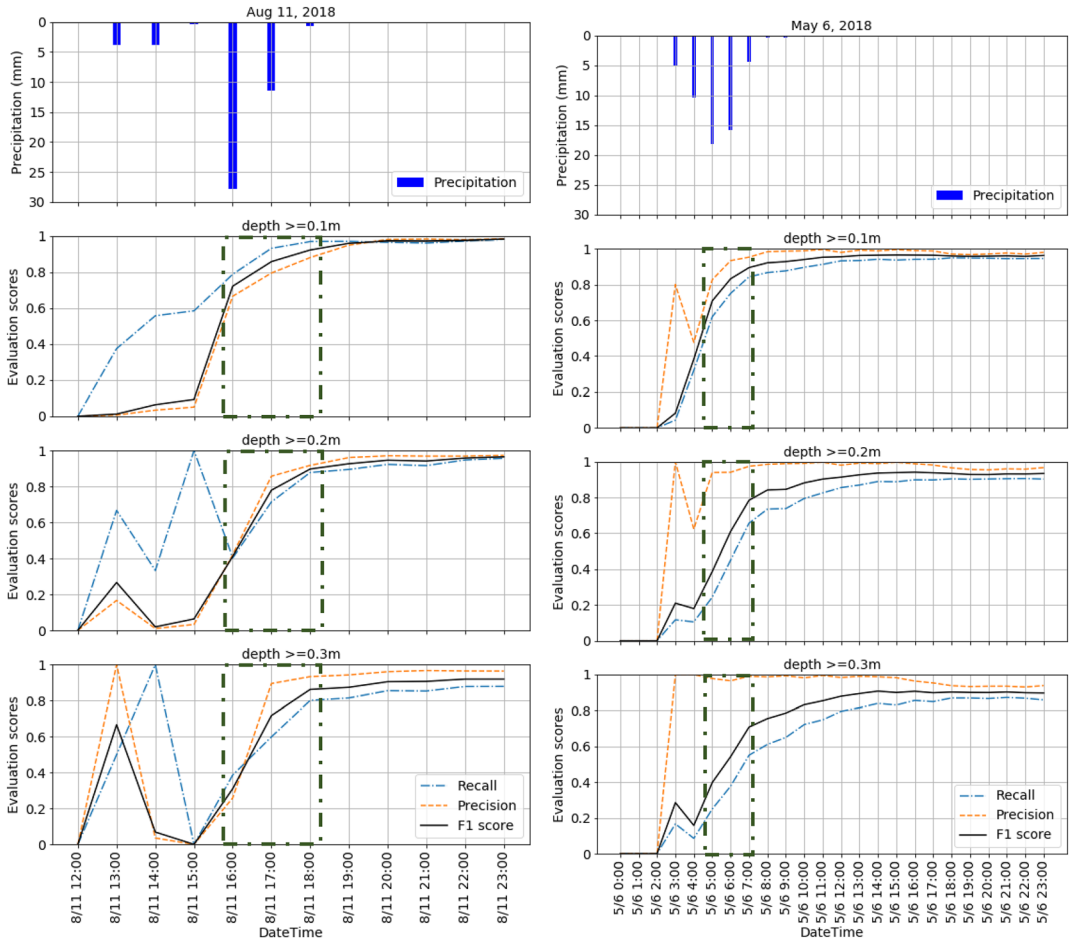


**Figure 11.** Time series plot of precision, recall, and F1 scores for threshold values 0.10, 0.20, and 0.30 m on 11 August 2018 and 6 May 2018 events. The dashed rectangle represents the 3 hr period with maximum volumes of water on the streets simulated by TUFLOW.

0.6 to 1.5 m by the TUFLOW model. Although the RF model also predicted high water depths, they were always smaller than the TUFLOW peak values, resulting in high RMSE. However, at 41 (e.g., labeled "C" in Figure 9) out of these 50 locations, RF Model 2 predicted water depths near or higher than the 0.30 m threshold, which is sufficient to make road closure decisions. RF Model 2 predicted peak water depths at the other nine locations were below 0.30 m.

At the remaining 12 locations with high RMSE, RF Model 2 overpredicted the peak water depth compared to the TUFLOW simulation. At these 12 locations (e.g., labeled "D" in Figure 9), RF Model 2 predicted peak water depth ranged between 0.50 and 0.75 m on the 11 August 2018 event, while the water depth from TUFLOW simulation was zero or smaller than 0.10 m. Analyzing the topographic features at these locations showed that TWI values were above 7.46. The original training data set had only 8% samples with TWI > 7.46, and 80% of these locations had TUFLOW-simulated peak water depth >0.10 m on 11 August 2018 event. On the contrary, 58% of the locations with TWI < 7.46 had TUFLOW-simulated peak water depth >0.10 m. As higher TWI indicates higher runoff accumulation, we suspected RF Model 2 overpredicted the flood depth due to high TWI values at these 12 locations.

### 3.3.3. Inundation Extent and Hazard Mapping

Figures 11 and 12 show the precision, recall, and F1 score time series throughout the test storm events for threshold values 0.10, 0.20, and 0.30 m of water depth, along with the hourly precipitation averaged across all rainfall stations in the study area. These time series reflect the performance of RF Model 2 throughout each storm event. Table 5 shows the average precision, recall, and F1 scores for the 3 hr with maximum

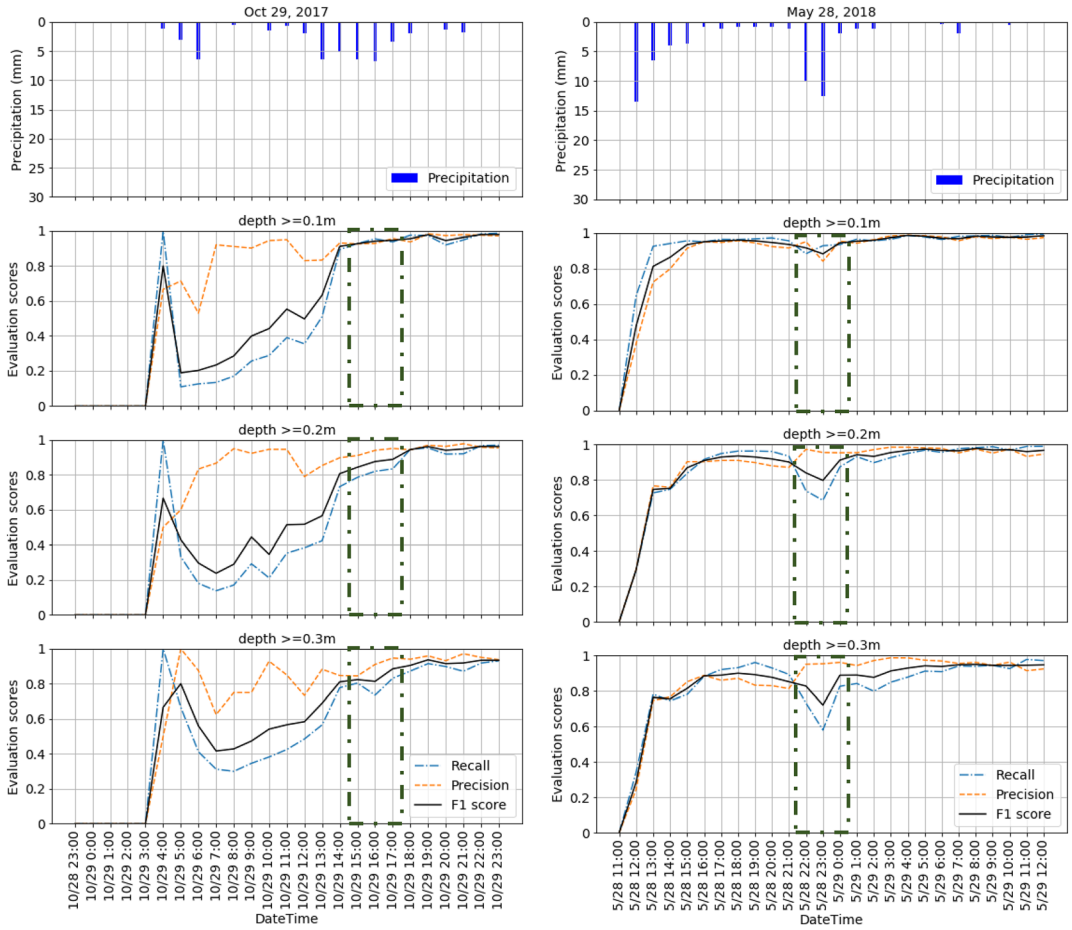


**Figure 12.** Time series plot of precision, recall, and F1 scores for threshold values 0.10, 0.20, and 0.30 m on 29 October 2017, and 28 May 2018 events. The dashed rectangle represents the 3 hr period with maximum volumes of water on the streets simulated by TUFLOW.

inundation volume according to TUFLOW simulation during the four test events. Figure 13 shows map-based views comparing water depth estimates from RF Model 2 and TUFLOW model for the 3 hr with maximum inundation volume during the 11 August 2018 event, along with flood reports from Waze.

Figure 11 shows that, during the initial hours of the storm events on 11 August 2018 and 6 May 2018, the precision, recall, and F1 scores were low. These hours corresponded to low water volume on the streets simulated by the TUFLOW model. However, with increasing hourly rainfall and cumulative rainfall in the previous hours, these scores gradually increased to satisfactory values. The F1 scores reached values above 0.85 for all the thresholds by the end of the 3 hr with maximum water volume on the 11 August 2018 event. Similarly, the 6 May 2018 event showed significant improvement in F1 scores during hours corresponding to maximum water volume. The 29 October 2017 event demonstrated a decrease in the F1 score curve during the reduced rainfall period. However, the average F1 scores were maintained to be >0.84 during the 3 hr of maximum water volume for the three threshold values. Although the F1 scores for the event on 28 May 2018 event declined during the period with maximum flood volume, the average value remained >0.80.

Table 5 shows that precision, recall, and F1 scores for water depths ≥0.1 m were high for the test storm events. Three out of the four test events had precision scores greater than 0.90, indicating more than 90% of the RF-predicted flooded road segments matched with TUFLOW-simulated flooded segments correctly. With increasing flood depth, the precision scores were consistently higher than 0.90 for these three events. The storm event on 11 August 2018 had the lowest precision score among the four storm events; however, it maintained a score higher than 0.70 for different flood depths. Figure 13 demonstrates that at 4:00 pm during

**Table 5**
*Precision, Recall, and F1 Score for the 3 hr With Maximum Flood Volume During the Test Storm Events*

| | Depth ≥ 0.1 m | | | Depth ≥ 0.2 m | | | Depth ≥ 0.3 m | | |
|---|---|---|---|---|---|---|---|---|---|
| Events | Recall | Precision | F1 score | Recall | Precision | F1 score | Recall | Precision | F1 score |
| 29 Oct 2017 | 0.94 | 0.94 | 0.94 | 0.81 | 0.93 | 0.87 | 0.79 | 0.90 | 0.84 |
| 6 May 2018 | 0.74 | 0.90 | 0.81 | 0.45 | 0.95 | 0.59 | 0.39 | 0.98 | 0.55 |
| 28 May 2018 | 0.92 | 0.92 | 0.92 | 0.77 | 0.96 | 0.85 | 0.71 | 0.96 | 0.81 |
| 11 Aug 2018 | 0.90 | 0.78 | 0.84 | 0.66 | 0.73 | 0.70 | 0.60 | 0.70 | 0.63 |

the storm event, which was also the hour with the highest flood volume, RF Model 2 overpredicted flooding in many locations, decreasing the precision value. The precision scores found in this study for water depths ≥0.1 m were within the range 0.19 to 0.97, which was also found in the study done by Bermúdez et al. (2018) and, therefore, could be considered acceptable scores.

In terms of recall, three out of the four test events had scores above 0.90 for water depths ≥0.1 m indicating more than 90% of the TUFLOW-simulated flooded streets defining the flood boundary were correctly predicted by RF Model 2. The storm event on 6 May 2018 had a recall score of 0.74, which was the lowest for water depth ≥0.1 m among the four test events. The recall scores found in this study for water depths ≥0.1 m were within the range 0.40 to 0.99, which are similar to those found in the study done by Bermúdez et al. (2018). For all four test events, recall decreased with increasing water depth. As discussed earlier in section 3.3.1, the data set used in this study was highly imbalanced. Oversampling was done on minority class with water depth ≥0.3 m to make the data set balanced, which notably improved overall precision and recall scores. However, the recall score for depth ≥0.3 m on the 6 May 2018 event was found to be

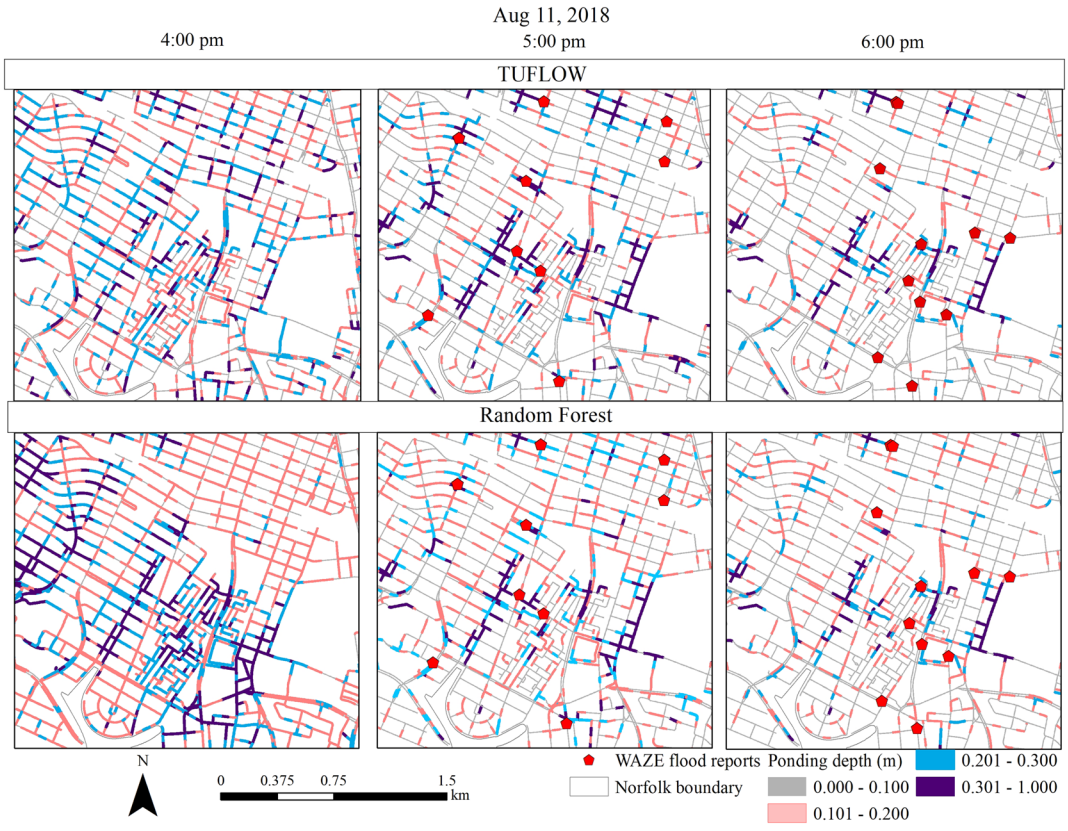


**Figure 13.** Comparison between inundation and hazard mapping by TUFLOW and RF Model 2 along with flood reports from Waze on 11 August 2018.

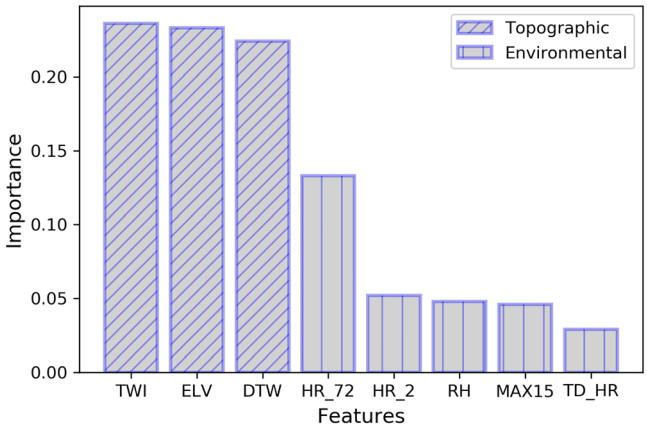


**Figure 14.** Importance of input features to RF Model 2.

0.39, which might indicate that the oversampling ratio of the minor class with water depth ≥0.3 m used in this study was not sufficient. In future research, oversampling the other minor class with water depth range 0.2–0.3 m and trying different oversampling ratios for the minority classes might help to improve recall scores further. Another approach to obtain a balanced data set is stratification, which could also be explored in future research. The performance of RF classification could also be further analyzed instead of using regression to increase precision and recall. Initial testing of RF classification with class_weight parameter as "balanced" in sklearn.RandomForestClassifier showed an increase in overall recall scores for water depths ≥0.2 m and ≥0.3 m, but a decrease for depths ≥0.1 m compared to RF regression.

Although recall decreased with increasing water depths, precision maintained high values regardless of water depth. This implies that the surrogate model underpredicted flooding, which might help to avoid unnecessary actions in nonflood locations but might miss some critical flooding impacts. Further exploring ways to balance the training data by oversampling minority class beyond what was tested in this study could further improve recall scores. However, due to the fundamental trade-off between precision and recall, increasing the recall values would likely result in a decrease in precision (Buckland & Gey, 1994; de Bruijn et al., 2017). Finding a balance between these two metrics would require understanding a decision maker's preference in either avoiding unnecessary actions or correctly identifying most of the flood locations, even if that may result in action taken in nonflooded locations.

### 3.3.4. Feature Importance

Feature importances from RF Model 2 are shown in Figure 14. The importance of each input feature is a relative measure calculated by RF based on how significantly the mean accuracy decreases if the feature was omitted. The three topographic features, TWI, elevation, and DTW, were found to be the most important features. TWI helped to explain pluvial flooding because it indicated the tendency of a pixel to receive and retain water from upstream. DTW represented the elevation difference between the land surface and nearest surface water features, so it likely accounted for flooding in tidally influenced areas. Because TWI and DTW were derived using the DEM, the impact to model results where elevation was excluded as a feature was explored. Doing this increased the average MAE by between 12% and 67% for the four test events, which supported including elevation in the RF model, despite its relation to TWI and DTW.

Among the environmental features, rainfall in the previous 72 and 2 hr were found to be more important than hourly rainfall, maximum 15 min rainfall, and tide level. This suggested that the RF surrogate gave more importance to antecedent moisture content and available capacity of storm water systems from prior rainfall occurrence compared to the more immediate rainfall for flood approximation. Investments in rainfall and soil moisture sensors within the city, therefore, could help to confirm real-time flooding impacts. Hourly tide level was the least important feature, although it was anticipated to play an important role given Norfolk's proximity to the coast and its low elevation. Surprisingly, omitting tide level from training reduced average MAE by between 0.13% (.00005 m) and 5.7% (.001 m) for the different test events. Because the improvement in MAE is a small number, and the hourly tide is one of the inputs for the TUFLOW model, this feature was not discarded from the analysis.

### 3.3.5. Computational Cost

Table 6 compares computational costs and run times of RF Model 2 to the TUFLOW model. The TUFLOW model simulated each storm event using two GPUs, which required approximately 4.5 to 6 hr of run time

**Table 6**
*Comparison Between the Machines Used and Computational Time for TUFLOW and RF Model 2*

| Model name | RAM | CPU | GPU info | Run time |
|---|---|---|---|---|
| TUFLOW | 64 GB | 4.4 GHz, 4 cores | Dual NVIDIA(R) GeForce(R) Titan X with 12GB GDDR5X each | 4.5–6 hr |
| RF surrogate | 16 GB | 3.6 GHz, 4 cores | — | Training: ~56 min<br>Testing: 6 s |

depending on the event duration. Using a CPU would most likely take more than 120 hr to simulate each event using the TUFLOW model (Morsy et al., 2018). In contrast, RF Model 2 required approximately 56 min to train using 16 events with a total of 375 hr of data, and it took only approximately 6 s to make predictions for a single event using the trained model. In addition, these times were for a computer with 16 GB of RAM and using a CPU rather than a GPU. This showed how the RF surrogate model would significantly reduce the computational cost for real-time flood management applications. It also sped up the simulation by a factor of around 3,000, which came with an arguably acceptable loss in accuracy, depending on the specifics of the decisions being made based on the modeling output. RF Model 2 could be trained using output from the physics-based model off-line before storms occur. Then, during storm events, RF Model 2 could be run in real time with a speedup from the physics-based model of 3,000 times.

### 3.4. Approach Limitations and Future Work

A major limitation of this study was the lack of observational water depth data on streets during storm events. Hence, it was presumed that the TUFLOW model was a ground truth, which might not always be the case as the Waze data suggested. A monitoring solution able to measure water depths on streets could be used to assess flood models better; however, such systems are rare, especially at a citywide scale. Although there are uncertainties and challenges associated with using crowdsourced data (Boutsis et al., 2016), crowdsourced data such as Waze are useful as an additional resource to validate street flood models, as shown in this study. Waze data are being used for other transportation-related assessments and decision making. For example, the U.S. Department of Transportation has been using Waze data for crash reporting and to enhance the predictive capability of crash models (Flynn et al., 2018).

Another major limitation of this study was the outdated underground pipe network data, which was last updated in 2000 by the city of Norfolk. As discussed in section 3.3.2, pipe network information was not available for some streets in the TUFLOW model, causing accumulation of water on the roadways. In addition, some of the pipes had higher downstream elevation compared to the upstream inlet, prohibiting water from draining out of the streets. Because a survey on the underground pipe network is costly, it makes the correction of the TUFLOW model at these problem spots difficult.

This study mainly focused on emulating the water depths on streets from the 1-D/2-D hydrodynamic model TUFLOW for real-time flood prediction. However, hydrodynamic models can simulate flooding both inside and outside urban streets. Future research could be performed to expand the flood prediction using the RF surrogate model outside the roads. An important feature to determine the magnitude of urban flooding is the available capacity of the drainage system. Due to the lack of data regarding initial conditions of the drainage infrustructure, rainfall data during the previous 2 and 72 hr were used as features to account for available storm water system capacity. It was assumed that the occurrence of rainfall in previous hours would reduce drainage system capacity, increasing the possibility of flooding with further rainfall. Also, in coastal cities, tide levels might interfere with the drainage capacity. Future studies should explore how drainage system capacity can be used directly as a feature in the RF model while considering impacts from both rainfall and tide levels.

While this work focused on urban coastal communities, the results could also be applied to other domains such as inland urban communities and low-relief rural communities that have complex hydrodynamics requiring computationally expensive 1-D/2-D physics-based models for street flood predictions. RF models would need to be retrained for these new application areas; however, this study did not explore the ability of an RF model to be applied to a new domain. Also, this study focused on an event-based sampling of training and testing data sets. Future studies might conduct a nonevent-based selection by choosing random time steps across all the events instead of partitioning training and testing by discrete rainfall events.

## 4. Conclusions

Machine learning models were developed using an RF algorithm to emulate results from the physics-based 1-D pipe/2-D overland model TUFLOW, which was built for a large portion of the coastal city of Norfolk, VA, USA. The RF surrogate models were trained to find patterns between topographic (TWI, DTW, and elevation) and environmental (rainfall and tide) features of roadways and water depths on streets simulated by the TUFLOW model for different storm events. Two different model training strategies were explored: (i)

training and testing using only the six most flood-prone street segments (referred to as RF Model 1) and (ii) training and testing across all of the 16,914 street segments in the nearly citywide data set (referred to as RF Model 2). The MAEs and RMSEs between water level predictions on street segments made by the RF surrogate models and TUFLOW simulations were calculated for both training strategies. In addition, precision, recall, and F1 score statistics were calculated to assess the performance of RF Model 2 in estimating the inundation extents simulated by the physics-based model.

Results showed a good predictive skill for both modeling scenarios. For RF Model 1, the averaged MAE and RMSE across the six flood-prone segments and four test events were 0.036 and 0.057 m, respectively, which increased by 0.015 and 0.023 m, respectively, for RF Model 2. For RF Model 2, the average and 90th percentile MAE varied between 0.012–0.039 m and 0.022–0.067 m, respectively, across the four test flooding events. The results of this comparison showed that the focused training approach produced a more accurate surrogate model for the flood-prone streets compared to the citywide RF Model 2, as expected, but RF Model 2 also performed well. Thus, we recommend an ensemble approach where RF models are built and trained for different collections of streets to produce the most accurate information to support decision makers for real-time management during flooding events. Furthermore, the results of the analysis showed that topographic features were more important compared to the environmental features in approximating water depths in the RF model. Among environmental features, cumulative rainfall over the previous 72 hr, representing antecedent moisture conditions, was the most important variable for predicting water depths on roads.

Due to a disproportionate number of streets with high water levels in the physics-based model compared to streets with little or no ponded water, the citywide RF Model 2 scenario tended to underestimate the water level. Efforts were made to balance the training data set, which did improve prediction results, although future research could further improve training strategies and address the imbalance. High precision (78–94%) and recall (74–94%) were obtained with a water depth threshold $\geq$0.10 m for all of the test events. However, increasing the threshold value to $\geq$0.30 m decreased recall (39–79%), while precision continued to be high (70–98%).

The major benefits of and motivations for using the RF surrogate models are their potential to reduce the computational time required while approximating the responses from the physics-based model with a sufficient level of accuracy to support decision makers. For real-time and forecast decision support, the computational demands of a detailed, 1-D/2-D physics-based model able to simulate street-scale water levels within an urban environment in real time is impractical if not impossible. RF model 2 had a 3,000 times speedup in run time compared to TUFLOW and did not require expensive hardware, like the dual GPUs used to run TUFLOW. This opens the door to real-time citywide flood forecasting capabilities at the street scale for urban communities without requiring access to powerful GPU workstations.

Lastly, an initial investigation of areas within the study domain where the RF surrogate model did not match simulations within the physics-based model suggested that some of these discrepancies could have resulted from problems in the physics-based model rather than problems in the RF model. For example, we found evidence to suggest missing storm water drainage infrastructure data for portions of the physics-based model domain that may explain unexpected water level predictions within these areas. Additional research, including data verification, field data collection, and model simulations, is needed to explore this finding more fully. Nevertheless, our research indicated that machine learning models like RF could be used as an approach to check physics-based models and identify potential problem spots that require additional attention and focus.

**Acknowledgments**

The work was supported by the National Science Foundation under Award Number CBET-1735587. The authors gratefully acknowledge the BMT for the TUFLOW HPC license. We would also like to acknowledge Waze and the city of Norfolk for providing roadway flood report data. We also wish to thank the Hampton Roads Sanitation District for access to their rainfall data. The resources used to perform this analysis, except for Waze data, are available on HydroShare (Zahura, 2019a, 2019b, 2019c). Waze data are available to government agencies via free Waze for Cities data-sharing program.

## References

Agranoff, D., Fernandez-Reyes, D., Papadopoulos, M. C., Rojas, S. A., Herbster, M., Loosemore, A., et al. (2006). Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. *Lancet*, *368*(9540), 1012–1021. https://doi.org/10.1016/S0140-6736(06)69342-2

Alexandrov, N., & Lewis, R. M. (2001). An overview of first-order model management for engineering optimization. *Optimization and Engineering*, *2*(4), 413–430. https://doi.org/10.1023/A:1016042505922

Alexandrov, N. M., Lewis, R. M., Gumbert, C. R., Green, L. L., & Newman, P. A. (2001). Approximation and model management in aerodynamic optimization with variable-fidelity models. *Journal of Aircraft*, *38*(6), 1093–1101. https://doi.org/10.2514/2.2877

AusRoads, (2008). Guide to Road Design, Part 5: Drainage Design.

Berkhahn, S., Fuchs, L., & Neuweiler, I. (2019). An ensemble neural network model for real-time prediction of urban floods. *Journal of Hydrology*, *575*, 743–754. https://doi.org/10.1016/j.jhydrol.2019.05.066

Bermúdez, M., Cea, L., & Puertas, J. (2019). A rapid flood inundation model for hazard mapping based on least squares support vector machine regression. *Journal of Flood Risk Management*, 1–14. https://doi.org/10.1111/jfr3.12522
Bermúdez, M., Ntegeka, V., & Wolfs, V. (2018). Development and comparison of two fast surrogate models for urban pluvial flood simulations. *Water Resources Management*, *32*(8), 2801–2815. https://doi.org/10.1007/s11269-018-1959-8
Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834
BMT WBM, (2016). TUFLOW User Manual.
Boutsis, I., Kalogeraki, V., & Guno, D. (2016). *Reliable crowdsourced event detection in smartcities*. Paper presented at 2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership With Global City Teams Challenge (GCTC), SCOPE—GCTC 2016, Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/SCOPE.2016.7515060
Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, *13*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x
Breen, K. H., White, J. D., & James, S. C. (2020). Are extreme soil moisture deficits captured by remotely sensed data retrievals? *Remote Sensing Letters*, *11*(8), 767–776. https://doi.org/10.1080/2150704X.2020.1766724
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L
Caviedes-voullième, D., García-navarro, P., & Murillo, J. (2012). Influence of mesh structure on 2D full shallow water equations and SCS curve number simulation of rainfall/runoff events. *Journal of Hydrology*, *448–449*, 39–59. https://doi.org/10.1016/j.jhydrol.2012.04.006
Chang, L. C., Shen, H. Y., Wang, Y. F., Huang, J. Y., & Lin, Y. T. (2010). Clustering-based hybrid inundation model for forecasting flood inundation depths. *Journal of Hydrology*, *385*(1–4), 257–268. https://doi.org/10.1016/j.jhydrol.2010.02.028
Chen, B., Harp, D. R., Pawar, R. J., Stauffer, P. H., Viswanathan, H. S., & Middleton, R. S. (2020). Frankenstein's ROMster: Avoiding pitfalls of reduced-order model development. *International Journal of Greenhouse Gas Control*, *93*, 102892. https://doi.org/10.1016/j.ijggc.2019.102892
City of Norfolk GIS Bureau, (2018a). Street Centerline|Norfolk Open GIS Data [WWW Document]. URL https://norfolkgisdata-orf.opendata.arcgis.com/datasets/street-centerline (accessed 10.1.19).
City of Norfolk GIS Bureau, (2018b). Interactive Norfolk|City of Norfolk, Virginia - Official Website [WWW Document]. URL https://www.norfolk.gov/1605/Interactive-Norfolk (accessed 4.6.20).
Danish Hydraulic Institute (DHI), (2017a). MIKE 11 Reference Manual, Danish Hydraulic Institute.
Danish Hydraulic Institute (DHI), (2017b). MIKE FLOOD User Manual.
de Bruijn, J., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. J. H. (2017). TAGGS: Grouping tweets to improve global geotagging for disaster response. *Natural Hazards and Earth System Sciences Discussions*, 1–22. https://doi.org/10.5194/nhess-2017-203
Deltares, (2018). Hydrodynamics, rainfall runoff and real time control. User Manual SOBEK.
Eggleston, J., & Pope, J. (2013). Land subsidence and relative sea-level rise in the southern Chesapeake Bay Region. US Geol. Surv. Circ. 1392.
Esri, (2020). ArcGIS Desktop|Desktop GIS Software Suite - Esri [WWW Document]. URL https://www.esri.com/en-us/arcgis/products/arcgis-desktop/overview (accessed 8.20.20).
Ezer, T., & Atkinson, L. P. (2014). Accelerated flooding along the U.S. East Coast: On the impact of sea-level rise, tides, storms, the Gulf Stream, and the North Atlantic Oscillations. *Earth's Future*, *2*, 362–382. https://doi.org/10.1002/2014EF000252
Fears, D. (2012). Built on sinking ground, Norfolk tries to hold back tide amid sea-level rise [WWW Document]. Washington Post. URL https://www.washingtonpost.com/national/health-science/built-on-sinking-ground-norfolk-tries-to-hold-back-tide-amid-sea-level-rise/2012/06/17/gJQADUsxjV_story.html (accessed 9.30.19).
Flood, J. F., & Cahoon, L. B. (2015). Risks to coastal wastewater collection systems from sea-level rise and climate change. *Journal of Coastal Research*, *27*, 652–660. https://doi.org/10.2307/41315838
Flynn, D. F. B., Gilmore, M. M., & Sudderth, E. A. (2018). Estimating traffic crash counts using crowdsourced data pilot analysis of 2017 Waze data and police accident reports in Maryland 2018 project summary. John A. Volpe National Transportation Systems Center (U.S.).
Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7
Jacobs, J. M., Cattaneo, L. R., Sweet, W., & Mansfield, T. (2018). Recent and future outlooks for nuisance flooding impacts on roadways on the U.S. east coast. *Transportation Research Record*, *2672*(2), 1–10. https://doi.org/10.1177/0361198118756366
James, S. C., Zhang, Y., & O'Donncha, F. (2018). A machine learning framework to forecast wave conditions. *Coastal Engineering*, *137*, 1–10. https://doi.org/10.1016/j.coastaleng.2018.03.004
Jhong, B. C., Wang, J. H., & Lin, G. F. (2017). An integrated two-stage support vector machine approach to forecast inundation maps during typhoons. *Journal of Hydrology*, *547*, 236–252. https://doi.org/10.1016/j.jhydrol.2017.01.057
Khu, S., Savic, D., Liu, Y., Madsen, H., & Science, C. (2004). *A fast evolutionary-based meta-modelling approach for the calibration of a rainfall-runoff model*. Paper presented at 2nd International Congress on Environmental Modelling and Software, Osnabrück, Germany.
Kleinosky, L. R., Yarnal, B., & Fisher, A. (2007). Vulnerability of Hampton roads, Virginia to Storm-Surge Flooding and Sea-Level Rise. *Natural Hazards*, *40*(1), 43–70. https://doi.org/10.1007/s11069-006-0004-z
Kulp, S. A., & Strauss, B. H. (2019). New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. *Nature Communications*, *10*(1), 4844. https://doi.org/10.1038/s41467-019-12808-z
Lane, T., & Brodley, C. E. (1997). An application of machine learning to anomaly detection. In *Proceedings of the 20th National Information Systems Security Conference* (Vol. 377, pp. 366–380). Baltimore, USA.
Leandro, J., Chen, A. S., Djordjevi, S., & Savi, D. A. (2009). Comparison of 1D/1D and 1D/2D coupled (sewer/surface) hydraulic models for urban flood simulation. *Journal of Hydraulic Engineering*, *135*, 495–504.
Lhomme, J., Bouvier, C., Mignot, E., & Paquier, A. (2006). One-dimensional GIS-based model compared to two-dimensional model in urban floods simulation. *Water Science and Technology*, *54*(6–7), 83–91. https://doi.org/10.2166/wst.2006.594
Lian, J. J., Xu, K., & Ma, C. (2015). Joint impact of rainfall and tidal level on flood risk in a coastal city with a complex river network: A case study of Fuzhou City, China. *Hydrology and Earth System Sciences*, *17*(2), 679–689. https://doi.org/10.5194/hess-17-679-2013
Liu, Y., & Pender, G. (2015). A flood inundation modelling using v-support vector machine regression model. *Engineering Applications of Artificial Intelligence*, *46*, 223–231. https://doi.org/10.1016/j.engappai.2015.09.014
Loos, M., & Elsenbeer, H. (2011). Topographic controls on overland flow generation in a forest—An ensemble tree approach. *Journal of Hydrology*, *409*(1–2), 94–103. https://doi.org/10.1016/j.jhydrol.2011.08.002

Madsen, J., & Langthjem, M. (2001). Multifidelity response surface approximations for the optimum design of diffuser flows. *Optimization and Engineering*, *2*(4), 453–468. https://doi.org/10.1023/A:1016046606831
Mark, O., Weesakul, S., Apirumanekul, C., Aroonnet, S. B., & Djordjević, S. (2004). Potential and limitations of 1D modelling of urban flooding. *Journal of Hydrology*, *299*(3–4), 284–299. https://doi.org/10.1016/S0022-1694(04)00373-7
McFee, B., & Lanckriet, G. (2010). *Metric learning to rank* (pp. 775–782). Paper presented at ICML 2010—Proceedings, 27th International Conference on Machine Learning.
Moftakhari, H. R., Aghakouchak, A., Sanders, B. F., & Matthew, R. A. (2017). Cumulative hazard: The case of nuisance flooding Earth's future. *Earth's Future*, *5*, 214–223. https://doi.org/10.1002/eft2.186
Morsy, M. M., Goodall, J. L., O'Neil, G. L., Sadler, J. M., Voce, D., Hassan, G., & Huxley, C. (2018). A cloud-based flood warning system for forecasting impacts to transportation infrastructure systems. *Environmental Modelling and Software*, *107*, 231–244. https://doi.org/10.1016/j.envsoft.2018.05.007
Muchoney, D., Borak, J., Chi, H., Friedl, M., Gopal, S., Hodges, J., et al. (2000). Application of the MODIS global supervised classification model to vegetation and land cover mapping of central america. *International Journal of Remote Sensing*, *21*(6–7), 1115–1138. https://doi.org/10.1080/014311600210100
Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.
Murphy, P., Ogilvie, J., Connor, K., & Arp, P. (2007). Mapping wetlands: A comparison of two different approaches for New Brunswick, Canada. *Wetlands*, *27*(4), 846–854. https://doi.org/10.1672/0277-5212(2007)27[846:MWACOT]2.0.CO;2
Murphy, P. N. C., Ogilvie, J., & Arp, P. (2009). Topographic modelling of soil moisture conditions: A comparison and verification of two models. *European Journal of Soil Science*, *60*(1), 94–109. https://doi.org/10.1111/j.1365-2389.2008.01094.x
NOAA, (2018a). Daily summaries station details: NORFOLK NAS, VA US, GHCND:USW00013750|Climate Data Online (CDO)|National Climatic Data Center (NCDC) [WWW Document]. URL https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00013737/detail (accessed 10.6.19).
NOAA, (2018b). Sewells Point - Station Home Page - NOAA Tides & Currents [WWW Document]. URL https://tidesandcurrents.noaa.gov/waterlevels.html?id=8638610 (accessed 10.1.19).
Ong, Y. S., Nair, P. B., & Keane, A. J. (2003). Evolutionary optimization of computationally expensive problems via surrogate modeling. *AIAA Journal*, *41*(4), 687–696. https://doi.org/10.2514/2.1999
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.
Peng, F., Feng, F., & Mccallum, A. (2004). *Chinese segmentation and new word detection using conditional random fields*. Paper presented at 20th International Conference on Computational Linguistics.
Pregnolato, M., Ford, A., Wilkinson, S. M., & Dawson, R. J. (2017). The impact of flooding on road transport: A depth-disruption function. *Transportation Research Part D: Transport and Environment*, *55*, 67–81. https://doi.org/10.1016/j.trd.2017.06.020
Prein, A. F., Liu, C., Ikeda, K., Trier, S. B., Rasmussen, R. M., Holland, G. J., & Clark, M. P. (2017). Increased rainfall volume from future convective storms in the US. *Nature Climate Change*, *7*(12), 880–884. https://doi.org/10.1038/s41558-017-0007-7
Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., & Kevin Tucker, P. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Science*, *41*(1), 1–28. https://doi.org/10.1016/j.paerosci.2005.02.001
Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*, W07401. https://doi.org/10.1029/2011WR011527
Resilient Cities, (2019). Resilient cities, resilient lives: Learning from the 100RC network. https://100resilientcities.org/capstone-report/
Rossman, L. A. (2004). *Storm water management model user manual. Version 5*. U.S. Environmental Protection Agency.
Sadler, J. M., Goodall, J. L., Morsy, M. M., & Spencer, K. (2018). Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. *Journal of Hydrology*, *559*, 43–55. https://doi.org/10.1016/j.jhydrol.2018.01.044
Scikit-learn Developers (2018a). 3.2.4.3.2. sklearn.ensemble. RandomForestRegressor—scikit-learn 0.21.3 documentation [WWW document]. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.Html (accessed 10.4.19).
Scikit-learn Developers (2018b). 1.11. Ensemble methods—scikit-learn 0.23.1 documentation [WWW document]. URL https://scikit-learn.org/stable/modules/ensemble.html (accessed 5.19.20).
Scikit-learn Developers (2018c). Model evaluation: Quantifying the quality of predictions [WWW Document]. URL https://scikit-learn.org/stable/modules/model_evaluation.html (accessed 10.6.19).
Shand, T. D., Cox, R. J., Blacka, M. J., & Smith, G. P. (2011). *Australian rainfall & runoff*.
Shen, Y. (2020). *Flood risk assessment and increased flood resilience for civil infrastructure in coastal regions under a changing climate*.
Shen, Y., Morsy, M. M., Huxley, C., Tahvildari, N., & Goodall, L. (2019). Flood risk assessment and increased resilience for coastal urban watersheds under the combined impact of storm tide and heavy rainfall. *Journal of Hydrology*, *579*, 124159. https://doi.org/10.1016/j.jhydrol.2019.124159
Simpson, T. W., Peplinski, J. D., Koch, P. N., & Allen, J. K. (2001). Metamodels for computer-based engineering design: Survey and recommendations. *Engineering Computations*, *17*(2), 129–150. https://doi.org/10.1007/PL00007198
Solomatine, D. P., & Torres, L. A. A. (1996). *Neural network approximation of a hydrodynamic model in optimizing reservoir operation* (pp. 201–206). Paper presented at International Conference on Hydroinformatics.
Sreetharan, M., Smirnov, D., Lawler, S., Schultz, M., Batten, B., Slover, K., et al. (2017). *Joint occurrence and probabilities of tides and rainfall*.
Suarez, P., Anderson, W., Mahal, V., & Lakshmanan, T. R. (2005). Impacts of flooding and climate change on urban transportation: A systemwide performance assessment of the Boston Metro Area. *Transportation Research Part D: Transport and Environment*, *10*(3), 231–244. https://doi.org/10.1016/j.trd.2005.04.007
Sweet, W. V., & Park, J. (2014). From the extreme to the mean: Acceleration and tipping points of coastal inundation from sea level rise Earth's future. *Earth's Future*, *2*, 579–600. https://doi.org/10.1002/2014EF000272
Syme, W. (2001). *TUFLOW-Two & onedimensional Unsteady FLOW software for rivers, estuaries and coastal waters*. Paper presented at IEAust Water Panel Seminar and Workshop on 2D Flood Modelling, Sydney.
U.S. Geological Survey, (2016). TNM download [WWW document]. Natl. Map. URL https://viewer.nationalmap.gov/basic/ (accessed 10.1.19).
US Department of Transportation, (2014). Mitigation strategies for design exceptions—Safety|Federal Highway Administration [WWW Document]. URL https://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm (accessed 10.5.19).
Vermeer, M., & Rahmstorf, S. (2009). Global sea level linked to global temperature. *Proceedings of the National Academy of Sciences*, *106*, 21,527–21,532.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wolfs, V., Meert, P., & Willems, P. (2015). Modular conceptual modelling approach and software for river hydraulic simulations. *Environmental Modelling and Software*, *71*, 60–77. https://doi.org/10.1016/j.envsoft.2015.05.010

Wolfs, V., & Willems, P. (2014). Development of discharge-stage curves affected by hysteresis using time varying models, model trees and neural networks. *Environmental Modelling and Software*, *55*, 107–119. https://doi.org/10.1016/j.envsoft.2014.01.021

Wolfs, V., & Willems, P. (2017). Modular conceptual modelling approach and software for sewer hydraulic computations. *Water Resources Management*, *31*(1), 283–298. https://doi.org/10.1007/s11269-016-1524-2

Yan, S., & Minsker, B. (2006). Optimal groundwater remediation design using an adaptive neural network genetic algorithm. *Water Resources Research*, *42*, W05407. https://doi.org/10.1029/2005WR004303

Yang, T., Asanjan, A. A., Welles, E., Gao, X., Sorooshian, S., & Liu, X. (2017). Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resources Research*, *5*, 2786–2812. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016). Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme Tiantian. *Water Resources Research*, *52*, 1626–1651. https://doi.org/10.1002/2015WR017394

Yin, J., Yu, D., Yin, Z., Liu, M., & He, Q. (2016). Evaluating the impact and risk of pluvial flash flood on intra-urban road network: A case study in the city center of Shanghai, China. *Journal of Hydrology*, *537*, 138–145. https://doi.org/10.1016/j.jhydrol.2016.03.037

Zahura, F. (2019a). Script for real-time street flood prediction model using machine learning, Norfolk, VA|CUAHSI HydroShare [WWW document]. URL https://www.hydroshare.org/resource/981253b3fbf5465fa11e0694c0015552/ (accessed 12.13.19).

Zahura, F. (2019b). Input data for real-time street flood prediction model using machine learning, Norfolk, VA|CUAHSI HydroShare [WWW document]. URL https://www.hydroshare.org/resource/47a45c3185074e0e8a668babc396b4f2/ (accessed 12.13.19).

Zahura, F. (2019c). Output from Random Forest surrogate model for real-time street flood prediction in Norfolk, VA,|CUAHSI HydroShare [WWW document]. URL https://www.hydroshare.org/resource/164c1be0da184b1c84fe8b4e9d595533/ (accessed 12.15.19).