Journal of Hydrology 559 (2018) 43–55

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest

J.M. Sadler [a], J.L. Goodall [a,*], M.M. Morsy [a,b], K. Spencer [c]

[a] Dept. of Civil and Environmental Engineering, Univ. of Virginia, 351 McCormick Rd., P.O. Box 400742, Charlottesville, VA 22904, United States
[b] Irrigation and Hydraulics Dept., Cairo University, P.O. Box 12211, Giza 12614, Egypt
[c] Deputy Resilience Officer, City Manager's Office of Resilience, City of Norfolk 501 Boush St., Norfolk, VA 23510, United States

ARTICLE INFO

Article history:
Received 18 October 2017
Received in revised form 20 December 2017
Accepted 19 January 2018

This manuscript was handled by G. Syme, Editor-in-Chief, with the assistance of Jesús Mateo-Lázaro, Associate Editor

Keywords:
Flood modeling
Data-driven models
Urban hydrology
Coastal flooding

ABSTRACT

Sea level rise has already caused more frequent and severe coastal flooding and this trend will likely continue. Flood prediction is an essential part of a coastal city's capacity to adapt to and mitigate this growing problem. Complex coastal urban hydrological systems however, do not always lend themselves easily to physically-based flood prediction approaches. This paper presents a method for using a data-driven approach to estimate flood severity in an urban coastal setting using crowd-sourced data, a non-traditional but growing data source, along with environmental observation data. Two data-driven models, Poisson regression and Random Forest regression, are trained to predict the number of flood reports per storm event as a proxy for flood severity, given extensive environmental data (i.e., rainfall, tide, groundwater table level, and wind conditions) as input. The method is demonstrated using data from Norfolk, Virginia USA from September 2010 to October 2016. Quality-controlled, crowd-sourced street flooding reports ranging from 1 to 159 per storm event for 45 storm events are used to train and evaluate the models. Random Forest performed better than Poisson regression at predicting the number of flood reports and had a lower false negative rate. From the Random Forest model, total cumulative rainfall was by far the most dominant input variable in predicting flood severity, followed by low tide and lower low tide. These methods serve as a first step toward using data-driven methods for spatially and temporally detailed coastal urban flood prediction.

1. Introduction

Flooding in low-lying, coastal cities has become more common in recent years due to climate change and relative sea level rise (Sweet and Park, 2014). In the coming decades, relative sea level is expected to rise substantially (Vermeer and Rahmstorf, 2009; Church et al., 2001), which will worsen the problem of flooding in coastal cities. Flooding in coastal cities is caused by large, life-threatening, high-return period events such as Hurricanes Harvey and Irma whose flooding recently affected coastal cities in Texas and Florida USA, respectively. Additionally, many coastal cities have low-relief terrain and low elevation making stormwater drainage problematic. This can make coastal cities susceptible to flooding from smaller, low-return period events such as severe thunderstorms. The long-term effects of legacy engineering decisions can further add to an urban city's flood risk (e.g., the use of non-engineered fill used to reclaim streams which causes higher than average subsidence rates (Turner, 2004)).

The ability to accurately predict flooding allows decision makers to proactively mitigate the effects of flooding (Zevenbergen et al., 2008), which is key to a city's resilience to natural hazards (Godschalk, 2003). Accurate flood prediction allows decision makers to maximize safety in the case of large events, and minimize infrastructure damage and social and economic disruption in the case of smaller events. Accurate flood prediction also allows cyber-physical (or smart) stormwater systems to perform optimally, further mitigating the effects of flooding (Kerkez et al., 2016).

Modeling and predicting flooding in urban coastal environments can be challenging. Urban coastal floods are influenced by a combination of different environmental, geographic, and human-related factors (Gallien et al., 2014). Environmental factors that contribute to coastal flooding include rainfall, wind, tide levels, and ground water table levels. Geographic factors such as elevation, soil properties, proximity to the coast, and the land

* Corresponding author.
E-mail address: goodall@virginia.edu (J.L. Goodall).

https://doi.org/10.1016/j.jhydrol.2018.01.044
0022-1694/

use and land cover of the drainage area can influence whether a given location experiences flooding.

In urban settings human-related factors including built stormwater infrastructure which is often underground, and the condition of that infrastructure, also play a role in the location and severity of flooding. For example, clogged stormwater inlets and undersized stormwater pipes can increase the chance and severity of flooding. High tidal levels can inundate stormwater outfalls rendering them ineffective at draining stormwater to the ocean, a condition which will become more frequent with sea level rise. The need to accurately represent such systems and their changing conditions further adds to the complexity of urban flood modeling.

Urban coastal flood events can be modeled using physically-based 1D (Mark et al., 2004) or 2D models (Mignot et al., 2006; Hunter et al., 2008; Bates et al., 2005; Smith et al., 2011; Gallien et al., 2014). However, the simplified representations of reality used in physically-based models can be a limitation given the combination of variables and their interactions, and the complexity of the physical environment. Two-dimensional, hydrodynamic models make fewer simplifications compared to 1D models, however, this comes at a larger computational cost (Leandro et al., 2009) which makes executing, and especially calibrating, a 2D model difficult (Caviedes-Voullième et al., 2012).

Another modeling approach shown to be effective in many fields (Yang et al., 2017a) including hydrology (Solomatine and Ostfeld, 2008) is data-driven modeling. Data-driven models detect patterns in the data to map model inputs to model outputs without attempting to simulate the physical processes (Solomatine and Ostfeld, 2008). Thus, the relationship between the inputs and outputs is not assumed, as in physically-based models, but learned. While physical processes are not directly simulated using data-driven models, understanding of physical processes usually influences the selection of input variables used to predict the output variable (Booker and Woods, 2014).

The recent increase in availability of earth observation data, coupled with advances in machine learning algorithms, have expanded the possibilities and use of data-driven modeling in hydrology. Machine learning algorithms have been used extensively in hydrology for applications such as predicting reservoir operations (Yang et al., 2016), soil mineral weathering (Povak et al., 2014), streamflow (Yang et al., 2017b; Solomatine and Xue, 2004; Wang et al., 2009), groundwater potential (Naghibi et al., 2017), and groundwater level (Sahoo et al., 2017). Data-driven and machine learning algorithms in flooding applications specifically have been used by Tehrany et al. (2013), Wang et al. (2015), and Tien Bui et al. (2016) who predicted areas susceptible to flooding, Adamovic et al. (2016), who modeled flash flooding on a regional scale, and Solomatine and Xue (2004), who predicted streamflow for flood forecasting. Despite the expanded use of data-driven models in hydrology, few studies have used data-driven methods to model flooding within coastal urban environments. The closest work may be the statistical analysis of tidal records in the United States to estimate the amount of time that coastal cities have experienced flooding in the past several decades and project flooding in the coming decades (Ezer and Atkinson, 2014; Sweet and Park, 2014; Moftakhari et al., 2015; Ray and Foster, 2016).

The objective of this study is to use data-driven modeling to predict flooding severity for a given storm in an urban coastal setting. Crowd-sourced flood reports recorded during flood events are used for model training and are considered a proxy variable for flood severity. Although a more objective measure of flood severity is preferred to the number of flood reports (e.g., flood inundation depth and duration throughout the study domain), often such data is not available. Relevant environmental data (rainfall, tide levels, water table level, wind speed and direction) will be used as inputs to the model.

A data-driven approach is appropriate for this application due to the complexity of modeling urban coastal flooding, as discussed above, which makes using a physical model difficult. This paper will investigate and compare two different data-driven models, Poisson regression and Random Forest regression. Poisson regression is a generalized linear model and was selected because it is commonly used to model rare events (D'Unger et al., 1998) and a flood report, while increasing in occurrence, can still be considered a rare event. Random Forest was selected due to its wide use as a machine learning algorithm in hydrology applications (Yang et al., 2016; Wang et al., 2015; Loos and Elsenbeer, 2011) and other fields (e.g.,Mutanga et al., 2012; Svetnik et al., 2003).

The data-driven approach will be applied in Norfolk, Virginia USA. Norfolk and the surrounding Hampton Roads region is one of the most vulnerable metropolitan centers to coastal flooding in the USA (Fears, 2012). Since 2010, the City has collected quality-controlled, crowd-sourced street flooding reports for 45 storm events. In this study, the two data-driven models, Poisson regression and Random Forest regression, will be trained to predict the number of street flood reports per storm event, given the rainfall, tidal, water table, and wind characteristics of the storm event. The models will be evaluated and compared using primarily the root mean squared error (RMSE) and mean absolute error (MAE) between the predicted number of street flood reports and the actual number of street flood reports.

This is a first step toward the use of data-driven approaches in urban coastal flood modeling. Additionally, although Gaitan et al. (2016) employed exploratory methods to glean information from open spatial data, weather data, and user reports, the use of crowd-sourced data in the training and evaluation of data-driven predictive models for urban flood modeling has not been demonstrated or discussed thoroughly in the literature. This is relevant currently as multiple platforms now exist for collecting crowd-sourced information regarding urban flooding (Le Coz et al., 2016) and it can be expected that, due to the nearly universal use of internet connected devices, crowd-sourced data will continue to grow in volume. The results of the modeling also shed light on the relative importance of different environmental factors in predicting coastal flooding, another subject that has been given little attention in previous literature regarding urban coastal flooding.

The remainder of this paper will proceed as follows. First, background will be given describing the study area, the model input and output data, and an introduction to the data-driven models used. Next, the methods are presented describing the preparation of the data for the models and how the data-driven models were applied and evaluated. The model results are then presented and discussed, and finally conclusions are given.

2. Study area, data, and model background

2.1. Study area and street flooding record

Norfolk, Virginia USA, shown in Fig. 1, is an ideal study area for this research considering its vulnerability to flooding, its economic and military importance, and the availability of quality-controlled crowd-sourced data regarding flood occurrences for the city. Norfolk is one of the most vulnerable cities to coastal flooding in the USA due largely to land subsidence rates causing Norfolk and the surrounding area to experience relative sea level rise at a rate faster than the global average (Kleinosky et al., 2006). As home to the largest terminal of the Port of Virginia, the 3rd most used port on the East Coast of the US (The Port of Virginia, 2016), Norfolk plays