Version of Record: https://www.sciencedirect.com/science/article/pii/S0198971522000989
Manuscript_ceed927488e61021d027535bbbaa764e

A Machine Learning Approach to Evaluate the Spatial Variability of New York City's 311 Street Flooding Complaints

Candace Agonafir[ab]*, Tarendra Lakhankar[b], Reza Khanbilvardi[ab], Nir Krakauer[ab], Dave Radell[c], Naresh Devineni[ab]*

a Dept. of Civil Engineering, The City University of New York (City College) New York, NY 10031, United States

b National Oceanic and Atmospheric Administration Center for Earth System Sciences & Remote Sensing Technologies (NOAA-CESSRST)

c National Weather Service, National Oceanic and Atmospheric Administration, US Department of Commerce, Upton, NY 11973

* Contact Authors: Candace Agonafir (cagonaf000@citymail.cuny.edu) and Naresh Devineni (ndevineni@ccny.cuny.edu)

Address:

Grove School of Engineering

160 Convent Avenue #T126

New York, NY 10031

Abstract

Urbanization, accompanied by the creation of roads, pavements, and sidewalks creates an environment where there is limited infiltration capacity, leaving metropolitan areas especially vulnerable during intense rain events. Furthermore, within an urban setting, there is spatial variability, as certain areas, owing to location, topography, land feature conditions, population and physical attributes or precipitation patterns, are more prone to flood damages. To detect neighborhoods with increased flood risk, crowdsourced data, which is the consolidation of eyewitness accounts, affords particular value. With an intent to understand how factors affect the spatial variability of street flooding, the Random Forest regression machine learning algorithm is employed, where the 311 street flooding reports of New York City (NYC) serve as the response, while the explanatory variables include topographic and land feature, physical and population dynamics, locational, infrastructural, and climatic influences. This study also analyzes socio-economic variables as predictors, as to allow for better insight into potential biases within the NYC 311 crowdsourced platform. It is found that catch basin complaints have overwhelmingly the greatest predictor importance, at 41%, almost sixfold higher than that of the second highest ranked predictor, slope, at 6.7%. Thus, NYC has an apparent issue with debris blocking the basins, and this may be remediated by increased cleaning efforts or public awareness to maintain clear streets, particularly during forecasted rain events. Furthermore, more than a third of the top predictors are land feature and topographical conditions, with building characteristics dominating the category. Often excluded in urban flood models, building effects, with a combined total importance of 11.7%, have greater significance than commonly considered flooding factors, such as percent impervious cover or elevation. Another major finding is the significance of the 'commuters who drive alone' variable, which alerts to the prospect of more reports being filed by those more affected by street flooding, as opposed to reflecting the actual occurrence of flooding (more reports being filed by those who drive on flooded roads versus those who do not). Overall, the leading contribution of this study is the identification of the top flooding factors in NYC, along with the presentation of their specific impacts towards street flooding variability among zip codes.

Highlights

- Catch basin complaints have the largest significance towards street flooding report variability.
- Building factors, often neglected in urban flood models, are found to have a greater effect on the spatial variability of street flooding than variables, such as elevation and percent of impervious cover.
- There is a degree of bias in the NYC 311 platform, especially in regards to commuters who drive to work.

1 Introduction

Perilous situations arise when urban flooding occurs. Posing a serious threat to life, rainwater, unable to enter the drainage network, ascends to considerable levels, overflowing the streets and sidewalks. Subsequently, individuals, unprepared for the flooding event, may drown in submerged basements or vehicles, become carried away by the waters, or endure fatal injuries by collapsed buildings and fallen trees. The devastation following Hurricane Ida (known as the post-tropical depression Ida) is a recent example of the human endangerment by urban flooding. The heavy downpour of the post-tropical depression resulted in 91 reported fatalities across nine states (Hanchey et al., 2021), including 13 New York City (NYC) area death (Plumer, 2021) s. In addition to this potential loss of life, urban flooding incurs substantial financial strain. During a flood event, widespread damage is sustained upon structures, train systems, electrical systems, and more, bringing forth extensive costs to repair. For instance, there was significant localized disruption of the NYC subway system and road transportation network during Ida; indeed, the weather disaster incurred one of highest recorded insurance losses in the U.S. at $36 billion (Aon, 2021). Moreover, the Federal Emergency Management Agency (FEMA) states that the combined urban flooding expenses for NYC and New Orleans metropolitan areas, over a 10-year period, totaled $10 billion (National Academies of Sciences, Engineering, 2019). Such high intensity and high total rainfall events are expected to become more frequent in a changing climate, and the juxtaposition of their space-time structure with the urban landscape and drainage systems (accounting for their reduced capacity due to blockages) determines the ultimate exposure of population and assets to flooding. Thus, due to the human and economic consequences of urban flooding, it is essential to identify areas of high flood risk as to allow for preventive measures.

Currently, there are models that forecast flash floods. In the United States, one of the most notable models is that by the National Weather Service (NWS). When there is a high intensity rainstorm or rainfall of sufficient duration that poses a flooding threat, the NWS will issue a flash flood watch or warning for the metropolitan area. Yet, the warning is based on observed heavy rainfall (NWS, 2022b), and it does not take into account land surface conditions or the drainage network. However, the NWS does offer a Flash Flood Guidance (FFG), which incorporates soil and streamflow conditions (NWS, 2022a). In addition, in a city such as NYC, which encompasses 800 square kilometers (United States Census Bureau, 2012), a warning system with more localized prediction will have greater utility. For instance, it may be difficult for all NYC basement apartment residents to vacate during a city-wide flash flood warning. However, if predicted at a finer spatial scale, for example, at the zip code level, the residents of the forewarned areas, perceiving a specific threat to their locations, may consider preventative measures, such as seeking shelter above ground. Moreover, in NYC, it has been shown that there is spatial variability in the occurrence of street flooding, where different regions may be more flood prone (Agonafir et al., 2021). Further, extreme rainfall, especially at the shorter durations, has considerable spatial variability (Hamidi et al., 2017). Thus, there remains a need to pinpoint problem areas within the urban domain.

Street flooding is influenced by a multitude of factors. First, there is the climatic factor. Precipitation, especially rainfall, is the major contributor, where an intense downpour of rain or rainfall for a long duration may overwhelm the drainage system, causing flooding (Sharif et al., 2006). Then, there are the topographical and land feature variables associated with flood risk, and these characteristics include the number of buildings, amount of impervious cover, slope, and elevation (Bruwier et al., 2020; Chang et al., 2015; Leandro et al., 2016; S.V. et al., 2015; X. Wang et al., 2019a). In addition, there are also engineering interventions, such as green roof installations, which may influence the ponding of water (Dietz, 2007), reduce peak runoff and impact the distribution of water resources (Asadieh & Krakauer, 2016). Finally, urban flooding research may examine infrastructural characteristics and population dynamics. For instance, the location and density of catch basins may impact water paths, and the concentration of people in an area may have an impact on the maintenance of the basins. Thus, there are many types of attributes within an urban environment which may affect the occurrence of flooding.

To evaluate the relative degree of flooding effect from each variable, the application of crowdsourced data has prospect. Crowdsourcing, a feature of social engagement that bridges the gap between researchers and data (Hedges & Dunn, 2018), often via an Internet platform, has been applied in numerous flood analyses and applications (Dede et al., 2019; Helmrich et al., 2021; Sadler et al., 2018; R. Q. Wang et al., 2018). For instance, utilizing Twitter, a flood detection platform in Indonesia, *PetaJakarta*, imports the flood-related tweets of residents to create real-time flood maps (See, 2019). Specifically in NYC, there is a crowdsourced platform, referred to as 311, where residents file reports of observed street flooding or infrastructural issues, and the locations of reports are recorded and available to the public (Minkoff, 2015). Furthermore, the NYC 311 database has been used in prior urban flood studies (Kelleher & McPhillips, 2020; Smith & Rodriguez, 2017; Agonafir et al., 2022). In Kelleher and McPhillips, 311 flooding reports were used to assess the impact of topographic wetness index and sink depth (Kelleher & McPhillips, 2020). Agonafir et al. examined the infrastructural predictors of NYC street flooding (Agonafir et al., 2021). If statistical learning tools, such as machine learning techniques, are utilized, then a relationship between each factor and the gathered crowdsourced accounts may be established, thereby providing illumination on the extent of the factor's impact.

Nonetheless, citizen generated information is potentially influenced by subsidiary motivations of the respondents. Some studies have shown crowdsourced projects to be biased, and despite being a platform open to the public, a small segment of the population may comprise a large portion of the responses (Basiri et al., 2019; Comber et al., 2016; Pak et al., 2017). For instance, in a Belgium-based platform, where residents report structural issues within their neighborhoods, Pak et al. found low-income groups were marginalized. As such, an exploration into the demographical differences may lend insight into the behavior and proclivities of participation (Dixon et al., 2021;

Moreno et al., 2015; Zhao & Zhu, 2014). Once participant motivation is discovered, the data may be curated to eliminate or minimize noise (Barbier et al., 2012). Therefore, analyzing potential outliers in crowdsourced data may optimize results and aid in the development of flood prediction models.

This paper presents an evaluation of the land and surface features, physical and population dynamics, climatic, and socio-demographic variables, via a Random Forest (RF) regression model, to discover the predictors of importance for NYC street flooding spatial variability. There are other machine learning algorithms which assess predictor effect. For instance, there is the highly regarded Extreme Gradient Boosting (XGBoost), an extension of the gradient descent methodology, which accommodates missing values (Rusdah & Murfi, 2020) and has an accuracy comparable with RF (Huang et al., 2020). However, XGBoost is not as resilient to noise, and consequently, it overfits (AlThuwaynee et al., 2021; Xu & Wang, 2019). RF, a Decision-Tree algorithm, is an ideal choice, as it also works well with missing values and datasets with a large number of predictor variables, of which only a fraction may actually be related to the response variable (Ali et al., 2012; Speiser et al., 2019). Moreover, RF functions effectively with outliers, and shows less overfitting than many algorithms (Liu et al., 2012; Rodriguez-Galiano et al., 2012).

This study hypothesizes that physical differences, such as precipitation pattern, percent impervious cover, slope, elevation, and the presence of buildings, affecting the natural processes of infiltration, have major contribution in street flood *occurrence*. To provide a holistic presentation, this paper also considers how the demographic (physical, financial, and behavioral) characteristics of the residents, affecting proclivity towards addressing concerns within a crowdsourced platform, has contribution towards street flood *reporting*. By the novel inclusion of the socio-economic variables, causes of potential bias are illustrated, and this allowance of relative importance comparisons between direct flooding factors and socio-economic variables give the findings more credence. Thus, serious consideration may be given to the flooding factors which prevail, despite the background of those reporting street flooding. In the analysis, total 311 street flooding reports, aggregated per zip code, are taken as the response variable. Physical and population features, precipitation variables, land feature and topographical conditions, locational and socio-demographic factors, serving as predictors, are prescreened by the RF model, where only the top 15 variables are elected. With the top 15 variables and total 311 catch basin reports serving as explanatory variables and street flooding reports serving as the response, 50 RF simulations are conducted, and the median relative importance for each predictor is then computed. With the presentation of these leading explanatory factors, an understanding into the spatial variability of NYC street flooding reports is achieved.

A purpose of this study is to extend the results of *Understanding New York City Street Flooding through 311 Complaints* (Agonafir et al., 2021), which had examined the infrastructural predictors of NYC street flooding. The prior analysis, utilizing a weekly time-series via negative binomial generalized regression, discovered spatial variabilities within NYC. Specifically, the frequency of street flooding complaints was found to vary per zip code; in addition, it was revealed that zip codes differed in climatic and infrastructural predictor significance. This study builds upon these findings by serializing the spatial units (zip codes) [as opposed to serializing the time unit], as to discover the relative importance of each factor in relation to total street flooding complaints. Furthermore, this paper delves into the unexplored aspects of Agonafir et al. by including socio-demographic variables, which may have influenced the crowdsourced data. By extending the conclusions of Agonafir et al., this study aims to achieve a holistic view of NYC street flooding, allowing for broader implications towards other metropolitan areas.

The paper is structured in the following manner. In Section 2, the study area, input data, and model background are described. The study area, NYC, is discussed, with a focus on the urban and economic characteristics. The socio-demographic, land feature and topographic, climatic, physical and population variables are detailed. In addition, a description of the NYC 311 crowdsourced platform is provided. The RF model is also briefly introduced. In Section 3, the methodology is outlined, including data processing. The specific details of the RF regressions are set forth, with a diagram depicting each model and the factors serving as inputs. In Section 4, the results are presented. Then, Section 5 proceeds with a discussion of the results and their implications. Lastly, Section 6 concludes with a summary of the findings and their unique contribution towards resolving issues within urban flood research.

2 Study Area, Data and Model Background

2.1 Study Area

Located along the northeastern coast of the United States, NYC, distinctly impervious, populous, and dense, manifests the urban metropolitan (*Impact of NYW Bonds*, n.d.; United States Census Bureau, 2012). Additionally, as it contributes the largest portion of gross domestic product (GDP), at approximately $1.8 trillion annually (Bureau of Economic Analysis, 2021), the economic dynamics within NYC may have overarching extent, nationally. Thus, due to its urban features and economic impact, NYC is chosen as an ideal study area to investigate urban flood factors.

Furthermore, in NYC, essential details, such as the locations and widths of stormwater inlet drains and digitized maps of the sewer network, are publicly unavailable. As such, flood modeling is challenging, and alternative methods of assessing problems within the infrastructure are desired. Therefore, this study, incorporating the infrastructural issues and components, has direct utility to the city.

2.2 Input Data

2.2.1 NYC 311 Platform

NYC 311 is a service which affords residents and visitors the opportunity to file reports concerning a wide-range of local problems, from noise complaints to sewer-related issues (City of New York, 2022a). The complaints may be registered via telephone or website. For researchers, the data is accessible via the NYC Open Data website: data.cityofnewyork.us. Available from January 1, 2010 to the present, each report includes a date and time and the latitude and longitude coordinates of the location where the issue has taken place. Two sewer-related complaints are of interest to this study: Street Flooding (SF) and Catch Basin (CB). SF complaints will illuminate and provide a workable metric for the occurrence of street flooding, and CB complaints provide insight into an infrastructural causal factor, as when catch basins are unable to receive rainwater, either due to blockage or malformation, surface water level increases on the streets. For SF, the complainant may report observed flooding or ponding on a street (City of New York, 2022e). For CB, the complainant may report issues with the catch basins, such as clogging or defective grates (City of New York, 2022b).

2.2.2 Radar Data

Stage IV data, at 4 km polar-stereographic grids, are available at the National Center for Atmospheric Research (NCAR)/Earth Observing Laboratory (EOL) website, where hourly, 6-hourly and 24-hourly analyses may be retrieved (Du, 2011). The data is a mosaic, comprised of radar and gauge estimates, thereby benefiting from the temporal and spatial resolutions of radar (Thorndahl et al., 2017) and the direct measurement capabilities of gauges (Serrano, 2010). Snow measurements are incorporated; however, due to instrumental error at some gauge locations, snow values may not be accurately reflected by the Multisensor Precipitation Estimates (MPE) algorithm (Du, 2011). Subsequently, the precipitation values of the Stage IV dataset are considered as rainfall estimates (Hamidi et al., 2017).

2.2.3 Socio-Demographic, Land, and Population Data

The socio-demographic data was taken from the NYC Geodatabase, released by Baruch College, and based upon the 2014-2018 American Community Survey (ACS) data and the 2010 census demographic data and ZIP Code Tabulation Areas (Baruch College, 2021). There were 121 socio-demographic variables, separated per zip code (over 174 NYC zip codes), with the following categories: Households by Type, Fertility, School Enrollment, Educational Attainment, Residence 1 Year Ago, U.S. Citizenship Status, Language Spoken at Home, Employment Status, Commuting to Work, Income and Benefits, Housing Occupancy, Housing Tenure, Housing Value, Mortgage Status, Gross Rent, Sex and Age, Race, Hispanic or Latino and Race, Citizen – Voting Age Population, Zip Code ID.

The land feature, topography, and population data were available in shapefiles, downloaded from NYC Open Data website: https://opendata.cityofnewyork.us/. NYC Open Data is a database provided by the City of New York.

2.3 Model Background

2.3.1 Random Forest

To measure the relative importance of each variable in an analysis, RF regression is effective. The technique has been used in multiple hydrological analyses (Loos & Elsenbeer, 2011; Z. Wang et al., 2015; Yang et al., 2016), and specifically, in flood studies (Albers et al., 2015; Lin et al., 2021). Introduced in 2001 by Leo Breiman, RF is a machine learning algorithm, suitable for handling large data sets (Breiman, 2001; Liaw & Wiener, 2002; Sadler et al., 2018). A bagged ensemble of prediction trees is trained to estimate predictor importance, with the tree learner being defined by setting the parameters to name-value pair arguments (MathWorks, 2022). The algorithm experiences a type of learning over the quantity of regression trees (Breiman, 2001). Then, the random forest predictor is determined by taking the average value over the number of grown trees (Liaw & Wiener, 2002). The algorithm provides the relative importance of the input variables.

2.3.2 Predictor Details

When considering the variables to input, factors affecting the hydrological processes necessary for the extraction of runoff are considered. Regarding the variety of the land surface, infiltration has significant effect on flooding. An important component of the hydrologic cycle, infiltration is the absorption of water by the soils during a rain event.

Impermeable materials, such as concrete, cement, brick, stone, and tile, where there is no infiltration capacity, leave water unable to be abstracted into the soil (S.V. et al., 2015). Therefore, the percent of impervious cover per zip code is included as a variable, as it decreases infiltration, thereby increasing runoff. A map displaying the average impervious cover in NYC, per zip code is shown in Fig 1a.

Aside from land surface, topographical factors, such as elevation and slope, also affect the behavior of runoff. It is reasoned that water flows along a slope; thus, in comparison to flat surfaces, water is less able to stand and rise to the significant levels (Rahmati et al., 2020). Indeed, it has been shown that a steeper slope leads to lower peaks in stored runoff volume and lower mean water depth (Bruwier et al., 2020). In regards to elevation, despite the possibility of increased precipitation at higher elevated areas (Novikov, 1981), the areas of low elevation surrounded by higher elevated areas are at greater flood susceptibility (Bado & Bationo, 2018; Ouma & Tateishi, 2014). It may be theorized that low areas are more flood prone, as they are at the bottom of a sloped surface, where the water, ultimately, is able to pond (X. Wang et al., 2019b). Thus, influencing the ponding of water during rainfall, the mean percent rise (slope) and mean elevation are inputs for the model. The maps of mean elevation and slope are shown in Fig 1b and Fig 1c, respectively.

Furthermore, urban features, specifically buildings, play a role in flooding. Buildings, including their respective elevations, have the added impact of changing the geometry and path of the natural flow (Chang et al., 2015; Leandro et al., 2016). In addition, the rooftops of buildings, considered impervious surfaces, contribute to greater amounts of effective rainfall. Thus, roofs, and their respective drainage network and gullies, should be considered for urban flood modeling (Chang et al., 2015; Leandro et al., 2016). It is also worth noting that some buildings have green roof installations, which offset the increase in runoff, allowing for infiltration. Highlighting this aspect, in a study by Dietz, green roof implementations had been found to abstract 63% of rainfall (Dietz, 2007). In NYC, where a green roof is defined as a layer of vegetation comprised of waterproofing, a root barrier, water retention and drainage, a growing medium, and plants, there are incentives and mandates to ensure their installations (City of New York, 2021a, City of New York, 2021b, City of New York, 2021c). For NYC, maps displaying the number of buildings per zip are shown in Fig 1d, the sum of the areas of each building footprint within each zip code are shown in Fig 1e, the sum of the areas of each building footprint within each zip code per zip code area are shown in Fig 1f.

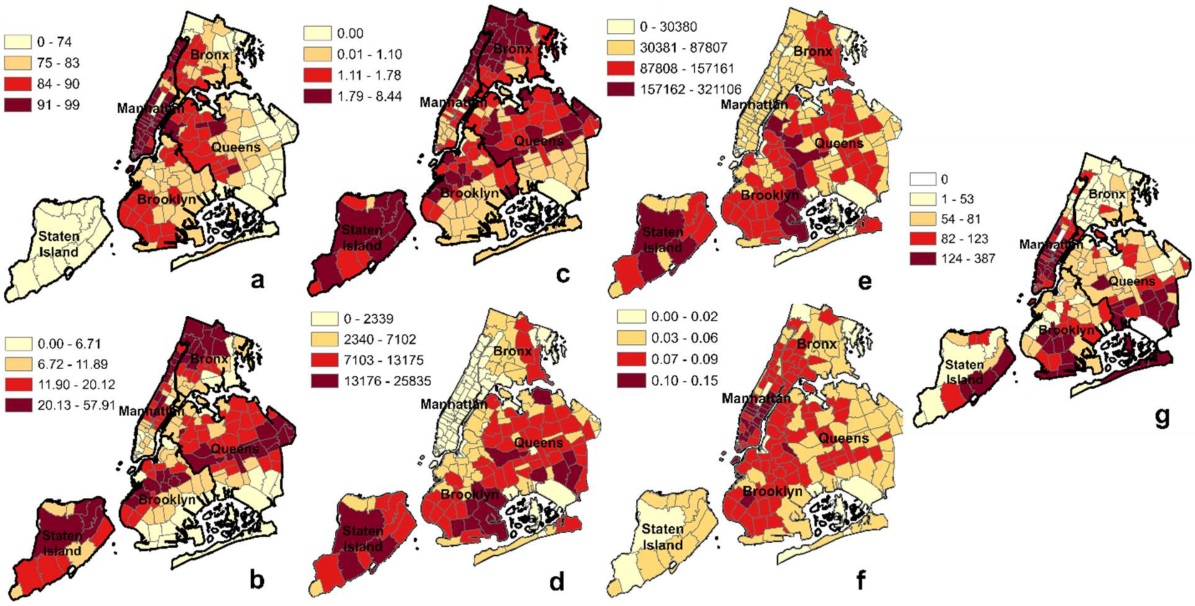

Fig 1 Maps displaying NYC land feature and topographical information and 311 SF frequency **Fig 1a** shows the percent impervious cover within each zip code **Fig 1b** shows the mean elevation in meters of each zip code **Fig 1c** shows the mean percent rise (slope) of each zip code **Fig 1d** shows the sum of the number of buildings within each zip code **Fig 1e** shows the sum of the building footprint areas in square meters within each zip code **Fig 1f** shows the sum of the building footprint areas per zip code area for each zip code **Fig 1g** shows the total SF complaints per zip code area in square kilometers

Increased Precipitation and Blocked Catch Basin Grates are distinguished by the NYC Department of Environmental Protection (DEP) as leading causes of street flooding in NYC (City of New York, 2022c). Precipitation, considered either rain, hail, or snow, is the primary driver of urban flooding. Specifically, rainfall is the major cause of flooding in urban cities, and in urban flooding model development, generally, total rainfall amount or rainfall intensity are used as inputs (Qin et al., 2013; Schmitt et al., 2004; Sharif et al., 2006). Concerning clogged catch basins, the basins are the inlets to the underground stormwater drains. During heavy rainfall, at times, debris, such as trash, construction waste, or leaves, are pushed on top of the catch basin grates, preventing rainwater from entering the sewer system. The water then ponds and rises to levels, considered as flooding. Thus, in accordance with the DEP and flood studies, the infrastructural issue of catch basin clogging, and the climatic cause of precipitation are considered.

Finally, as a measure to detect skewing of the 311 sewer-related reports, where background characteristics of the residents may affect inclinations to report, socio-demographical variables are included in the model. These factors do not physically influence street flooding; thus, they serve as an investigative technique to detect the accuracy of the 311 reports. For instance, two zip codes may have the same street flooding magnitude; yet, the zip code with the higher demographical bias may have more reports. Hence, if a socio-demographic variable is selected as a predictor, then in the areas where the particular variable trends, a greater frequency of SF reporting may not actually reflect a greater occurrence.

3 Data Processing and Methodology

The 311 SF and CB complaints were acquired from the NYC Open Data website. The reports from January 1, 2010 through December 31, 2019 were employed. The data was then geo-aggregated to the zip code level, with 174 zip codes being used for analysis. A measure, processing for uniqueness, was taken to evaluate whether a complainant was reporting more than once daily. Using the Distinct function, provided by R, the latitude and longitude coordinates of each complaint was examined. Of the raw 311 data, over the ten-year period, it was determined that 82,191 of the 85,607 CB reports (96.0%) and 25,378 of the 25,574 SF (99.2%) reports were unique.

Regarding the precipitation data, hourly totals from January 1, 2010 through December 31, 2019, were ordered from the EOL database. To analyze at the zip code level, radar points within the NYC boundary were extracted, and the Spatial Join, an ArcGIS analysis tool, was employed. By the method, a zip code was assigned to the radar point closest to its centroid. After the geoprocessing, there were a total of 40 radar points in NYC, and by applying the inverse distance weighting method, precipitation values were disaggregated to the 174 zip codes of this study. Firstly, concerning the short duration rainfall intensity variables, the mean hourly precipitation amounts of the non-zero values were calculated per zip code (mm/hour) and designated as NZMN (non-zero rainfall mean); in addition, of the non-zero values for the hourly data, the standard deviation, skewness, and kurtosis were determined and signified as NZSD (non-zero rainfall standard deviation), NZSW (non-zero rainfall skewness), and NZKT (non-zero rainfall kurtosis), respectively. Secondly, concerning the longer duration rainfall, daily totals were examined. The 95^{th} percentile values of the daily totals per zip code were computed and represented as PERC; also, from the daily totals, the mean and max length of the wet spell days were determined and represented by the parameters, MNWTS (mean wet spell length) and MXWTS (maximum wet spell length), respectively. Therefore, for the precipitation variables, hourly rainfall intensity and daily total statistics were utilized in the RF models.

Zip code, elevation points, impervious cover, number of buildings, building footprints, DEP green roof infrastructure, number of catch basin and borough shapefiles were downloaded from NYC Open Data and processed via ArcGIS Pro. The percent of impervious cover, population and area per zip code were provided within the Zip code shapefile. Mean elevation, mean slope (percent rise), and the centroid (x and y coordinates) per zip code were calculated with the utilization of ArcGIS Pro calculation tools. For building footprints, the area per footprint was calculated and the sum of the areas of each building footprint within each zip code was determined. Similarly, the area of the green roof installations, as listed within the DEP, was calculated with the sum for each zip code determined. The variables derived from the above-described processes are the following: mean percent rise (SLPE), mean elevation (ELEV), total area of green infrastructure (GREEN), catch basins per unit area (CBPA), population (POP), x coordinate of the centroid (XCOR), y coordinate of the centroid (YCOR), zip code area (AREA), population density (PPDN), percent of impervious cover (IMPV), number of buildings (BLD), the sum of the building footprints (FP), the sum of the building footprints per unit area (FPBD).

There were 121 socio-demographic variables per zip code provided, with categories such as educational attainment, household type, housing ownership profiles, sex, age, race, commuter status and income. The complete list of socio-demographic variables is shown in Appendix A.

For the determination of variable importance, the RF regression model was employed. MATLAB R2021a was utilized to perform the analysis. First, the tree learner was defined:

- All predictor variables were set to be used at each node.
- The predictor-selection technique was set to the interaction test, as it is the recommended method for analysis when the objective is determining predictor importance (Loh, 2004). In addition, it accommodates the possibility of local interactions between predictor variables during split selection (Loh, 2004).
- Surrogate splits were specified as to aid accuracy.

Once the template tree was established, a bagged regression ensemble model was created by the following inputs:

- The name-value pair argument was set to bootstrap aggregation.

- A bagged ensemble of 500 prediction trees were specified.

Out-of-Bag predictions (OOB) were then determined, and the explained variance, R^2, was calculated by the correlation between observed and predicted values of the response variable. Lastly, the *oobPermutedPredictorImportance* function was used, which provides Out-of-Bag, Predictor Importance Estimates by Permutation (impOOB). The impOOB values were also normalized as to scale the predictor importance value from 0 to 1. A more detailed description of the methodology (as implemented in the MATLAB R2021a Statistics and Machine Learning Toolbox) is provided in Appendix B.

In this study, there were multiple processes. As a preliminary step, before the set-up of the models, all variables were run via the Random Forest regression simulations (total of 142 variables), and it was seen that CB dominated the predictors, such that CB represented a median 22% of the relative importance, and the other variables represented a median 3% or lower relative importance, each. Thus, in order to clearly evaluate the effect of the other variables, it was decided that there would be two separate models, Model 1, featuring only the topographic, land feature, physical and population dynamics, and locational elements, and Model 2, where the infrastructural variable of CB would be included. Henceforth, for Model 1, there was a prescreening process, where SF reports served as the response variable, and the predictor variables were the land feature and socio-demographic variables (total of 141 variables). The RF regression simulation was run 50 times, and the median of the predictor importance values were determined. The R^2 is 0.58, and the results are shown in Appendix Table C.1. The purpose of the prescreening process was to allow the machine learning algorithm to filter the important variables. This prescreening procedure was implemented as a more suitable alternative than allowing a selection of variables by expert opinion. It was opted for the top 15 importance variables, as it serves as a tradeoff between too few and too many explanatory variables. The additional importance after the top 15 is less that 1%, and the total importance of the top 15 variables was in excess of 40%. Therefore, for the Model 1 results, the RF simulations were run again, but only with the 15 predictors shown to have highest importance. The median R^2 were also calculated from the 50 simulations of those 15 predictors. Next, as it was shown that CB reports influence SF reports in Agonafir et al., Model 2 repeats the process, with CB reports added as a predictor variable, along with the top 15 predictors of the initial RF analysis (i.e., from Model 1). Additionally, to gain additional insight into how the variables affect the crowd-sourced data, Model 3 was developed, where RF regression simulations were conducted with CB serving as the response variable and the explanatory variables being the same 141 variables as in the original regression. Again, to reduce noise, the RF simulations were repeated, but with the highest ranked 15 predictors. Model 3 serves as a background information about CB, an important SF determinant, and the results are listed in Appendix C. All models are depicted in Fig 2.

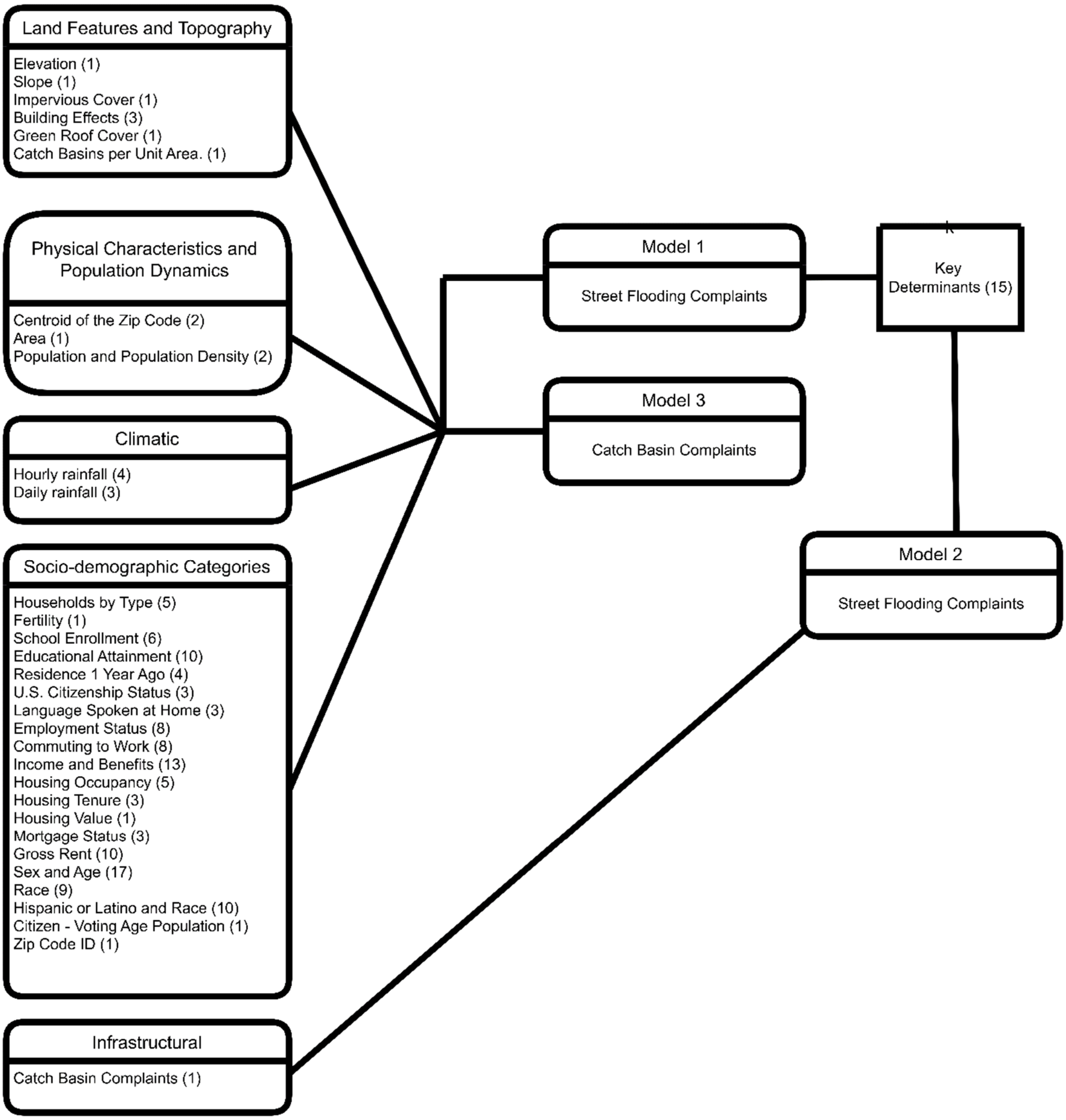

Fig 2 A diagram depicting the three RF regression models and the input variable types. In parenthesis are the number of variables within each variable category.

4 Results

4.1 Model 1: SF and Predictor Importance of the Land Features and Socio-demographic Variables

The top 15 predictors are the following: Commuting: drove alone (COM02), YCOR, BLD, AREA, SLPE, Employment status: armed forces (EMP06), Mortgage Status: mortgage (MORT02), FP, Housing tenure: Owner (HTEN02), ELEV, PPDN, IMPV, Residence 1 Year Prior: Abroad (RES04), FPBD, and XCOR. Running 50 simulations of the top 15 predictors only, the median R^2 is found to be 0.63, and box plots of the variables of each simulation are shown in Fig 3 and listed in Table 1.

Abbreviation	Variable	Percent Importance
COM02	COMMUTING TO WORK - Workers 16 years and over - Car, truck, or van -- drove alone	17.79
YCOR	Centroid of y coordinate	10.38
BLD	Number of buildings	9.43
AREA	Area	8.62
SLPE	Slope - mean percent rise	7.61
EMP06	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Armed Forces	6.58
MORT02	MORTGAGE STATUS - Owner-occupied units - Housing units with a mortgage	6.12
FP	Sum of the building footprints	6.10
HTEN02	HOUSING TENURE - Occupied housing units - Owner-occupied	5.13
ELEV	Mean elevation	5.10
PPDN	Population Density	4.04
IMPV	Percent of Impervious Cover	3.68
RES04	RESIDENCE 1 YEAR AGO - Population 1 year and over - Abroad	3.32
FPBD	Sum of the building footprints per unit area	3.22
XCOR	Centroid of x coordinate	3.11
		R^2 is 0.63

Table 1 The median R^2 and relative importance values, resulting from 50 simulations of the RF regression for only the top 15 ranked predictors, with SF serving as the response variable (Model 1)

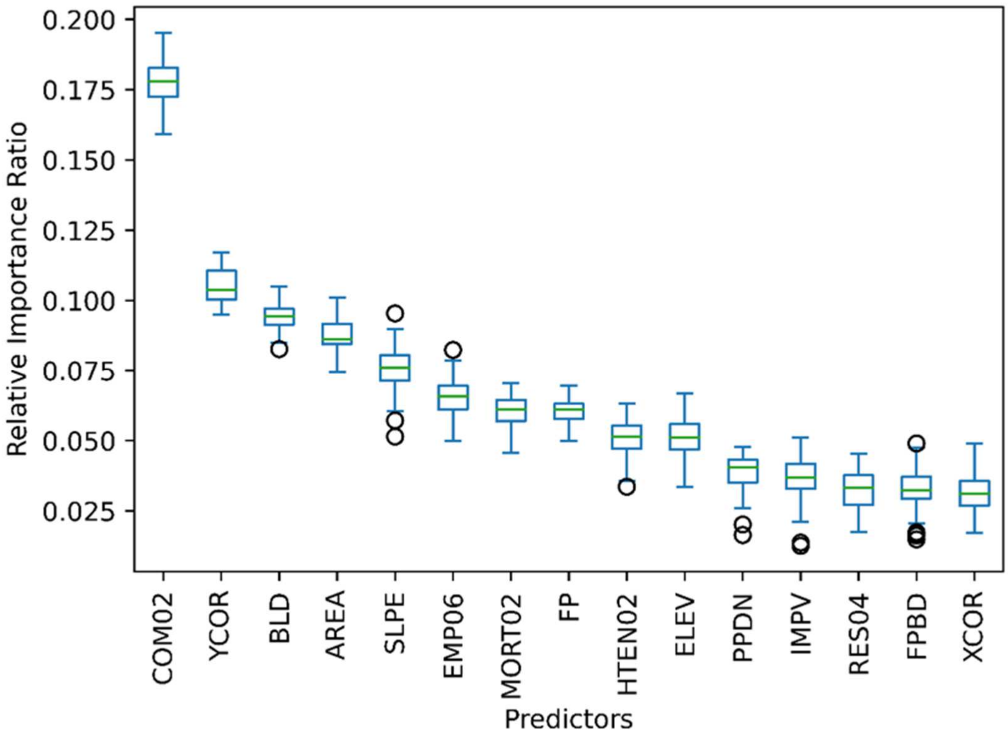

Fig 3 Box plots of the 50 RF simulations of the top 15 ranked variables only, with SF serving as the response. These 15 variables explain up to 63% of the spatial variability (The R^2 is 0.63). The expanded version of the acronyms is shown in Table 1.

Of the top categories, five socio-demographic categories were shown: Commuting to Work, Employment Status, Mortgage Status, Housing Tenure, and Residence 1 Year Ago. The Commuting to Work category includes remote workers, drivers, carpoolers, and those who take public transportation; Employment Status differentiates those who are either employed in armed forces or civilian forces or unemployed; Mortgage Status designates between those who have a mortgage on their properties and those who do not; Housing Tenure separates those who rent and those who own homes; lastly, the status of the Residence 1 Year Ago quantifies the population who lived in the same house, different house, or abroad the year before.

4.2 Model 2: SF and The Top 15 Land Feature and Socio-demographic Predictors including CB reports as a Predictor

Given that CB is an important variable considered by NYC local planners (City of New York, 2022c), total CB reports per zip code were added as a predictor, along with the resultant 15 top predictors of Model 1. With SF serving as the response, the median R^2 of the 50 RF simulation runs increases from 0.63 to 0.71. The RF results show CB as the most important predictor, at 41.13%, and it dominates the ratio of importance. The second most contributing predictor is SLPE, at 6.73%, and of the RF analyses, this represents the largest delta difference, at 34.40%. When only the top 15 predictors were run against SF in Model 1, slope previously obtained a 7.61% relative importance. Furthermore, once CB was added, other predictors experienced decreases in importance, when compared to Model 1 results. These include COM02, YCOR, BLD, and AREA, which decreased from 17.79% to 6.56%, 10.38% to 6.19%, 9.43% to 3.47%; and 8.62% to 5.12%, respectively. Box plots of the variables of each simulation are shown in Fig 4 and listed in Table 2. As an additional illustration, scatter plots for each predictor of Model 2 are given in Fig 5.

Abbreviation	Variable	Percent Importance
CB	Total catch basin complaints	41.13
SLPE	Slope - mean percent rise	6.73
COM02	COMMUTING TO WORK - Workers 16 years and over - Car, truck, or van -- drove alone	6.56
YCOR	Centroid of y coordinate (latitude)	6.19
AREA	Area	5.12
FPBD	Sum of the building footprints per unit area	4.79
EMP06	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Armed Forces	4.53
IMPV	Percent of Impervious Cover	4.37
BLD	Number of buildings	3.47
RES04	RESIDENCE 1 YEAR AGO - Population 1 year and over - Abroad	3.44

FP	Sum of the building footprints	3.43
MORT02	MORTGAGE STATUS - Owner-occupied units - Housing units with a mortgage	3.40
PPDN	Population Density	3.29
ELEV	Mean elevation	2.12
HTEN02	HOUSING TENURE - Occupied housing units - Owner-occupied	1.52
XCOR	Centroid of x coordinate (longitude)	0.02
		R^2 is 0.71

Table 2 The median R^2 and relative importance values, resulting from 50 simulations of the RF regression for only the top 15 ranked predictors and CB, with SF serving as the response variable (Model 2)

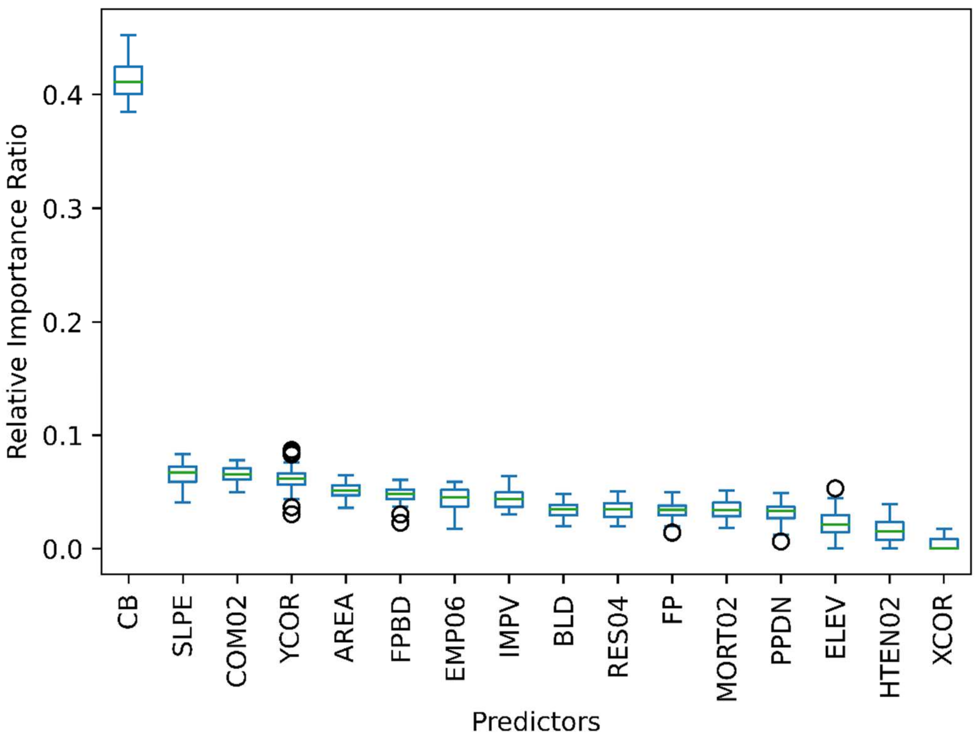

Fig 4 Box plots of the 50 RF simulations of the top 15 ranked variables and CB, with SF serving as the response. These 16 variables explain up to 73% of the spatial variability (The R^2 is 0.73). The expanded version of the acronyms is shown in Table 1.

5 Discussion

As demonstrated by previous studies, there is spatial variability in SF reports within NYC. Compiling the factors in each zip code as predictors and running RF regression simulations against the total SF complaints per zip code, over the course of 174 zip codes, enables insight into which factors have explanatory power for the local differences in SF reporting. Moreover, the RF regressions also provide percent importance. The valuations of each factor's effect allow a perception into an area's vulnerability based on the physical characteristics it contains; and, with this knowledge, urban risk assessment may be facilitated. Another advantage of the RF method is the assessment of predictor effect despite non-linear relationships. As seen in the scatterplots of Figure 5, not all the predictors of the study have linear relationships with the response; thus, a linear regression or other parametric modeling techniques would not be appropriate here. It is important to note, however, that, as the analysis is conducted over each zip code, the RF model results indicate predictors' importance in regards to spatial variability; thus, while a factor may be a significant contributor to street flooding (e.g., the temporal distribution of rainfall and intensity), if the values do not vary greatly per zip code, it will be designated with lower relevance.

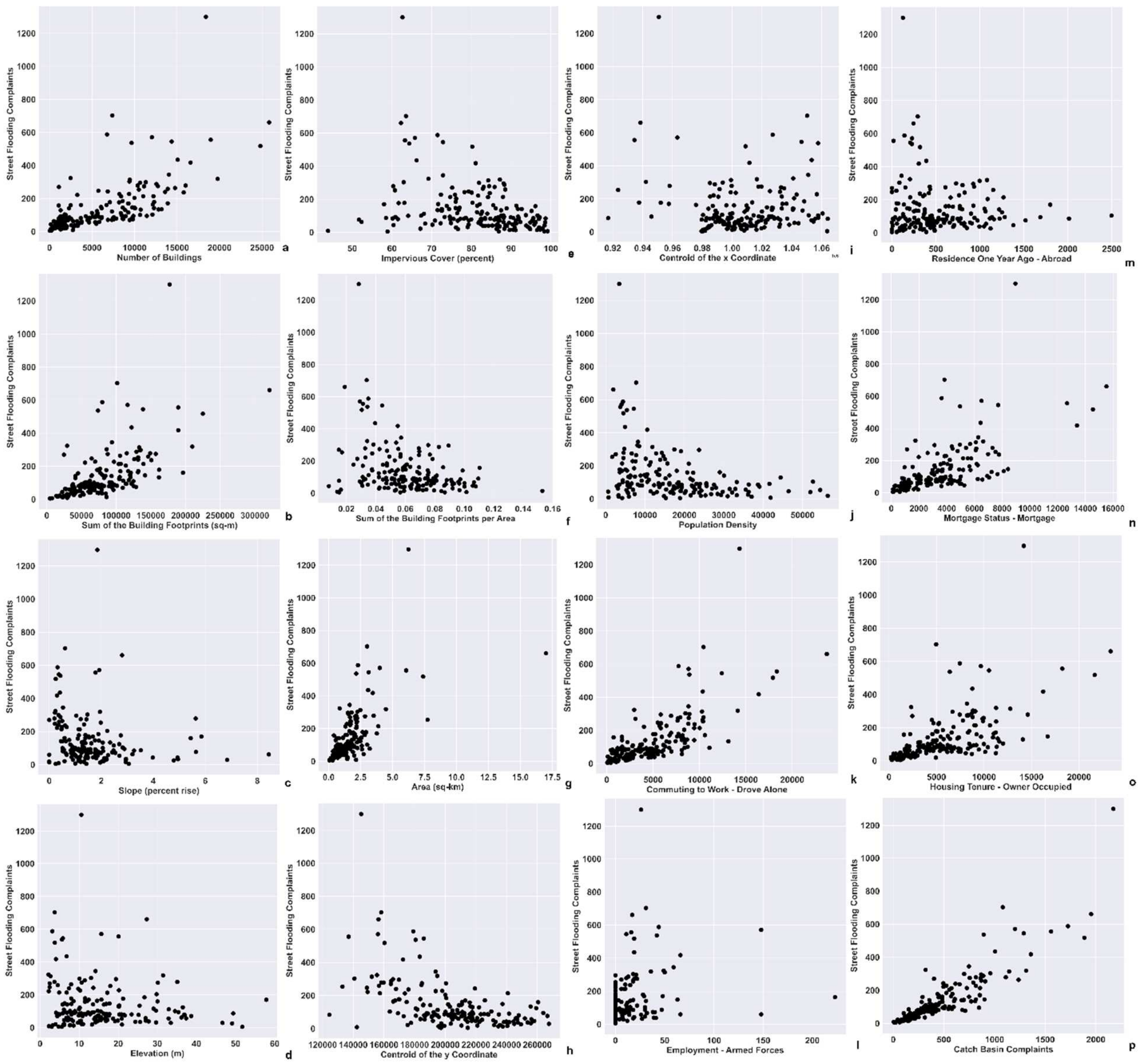

Fig 5 Scatter plots of the Model 2 predictors. Each dot represents a zip code. For each plot, SF complaints are on the y-axis, and the predictor is on the x-axis. The predictors are shown in the following: **Fig 5a** is BLD **Fig 5b** is FP **Fig 5c** is SLPE **Fig 5d** is ELEV **Fig 5e** is IMPV **Fig 5f** is FPBD **Fig 5g** is AREA **Fig 5h** is YCOR **Fig 5i** is XCOR **Fig 5j** is PPDN **Fig 5k** is COM02 **Fig 5l** is EMP05 **Fig5m** is RES04 **Fig 5n** is MORT02 **Fig 5o** is HTEN02 **Fig 5p** is CB.

5.1 Land Feature and Topographical Factors

The results of this study demonstrate that land features and topography have impact on the reporting of SF. In Model 1, over a third of the top 15 predictors are feature and surface characteristics. These include BLD (Number of buildings), FP (Sum of the building footprints), SLPE (Slope – mean percent rise), ELEV (Elevation), IMPV (Percent of impervious cover), and FPBD (Sum of the building footprints per square area). The totaled percent importance of the land feature and topographical factors is 35.14%. It is noteworthy that, within this category, the building factors, BLD, FP, and FPBD, combined, comprise 18.75% percent importance, which is more than the combined total of SLPE, ELEV, and IMPV at 16.39%. While many urban flood studies and models include topographical aspects, such as slope and elevation, building factors may be neglected (Lin et al., 2021). Indeed, the digital elevation model, which includes slope and elevation and often excludes buildings, serves as the basis for a vast quantity of urban flood models, especially for the 1D (one-dimensional) models (Bulti & Abebe, 2020; el Kadi Abderrezzak et al., 2009; Sharif et al., 2006). Additionally, given that NYC is currently exploring and implementing green roofs mandates (City of New York, 2022f), the building factors finding is valuable. A further inference is that all metropolitans may not be treated equally. For instance, in modeling a city such as NYC, where there is a marked presence of buildings, the inputs may be weighted differently than when modeling a major city, where, perhaps, varying elevations is the distinct characteristic. This study brings to light the possibility that not all flooding factors are universally significant to the same extent. Therefore, the distinguishment of building effects by the RF model strengthens the importance of their inclusion in future research.

5.2 Physical Characteristics and Population Dynamics

The physical location of the zip code, the zip code area, and the population density are also factors found to have effect on regional differences within SF reporting. For Model 1, AREA (area), YCOR (centroid of y coordinate), PPDN (population density), and XCOR (centroid of the x coordinate) are amongst the top 15 explanatory factors. The area of the zip code affects the quantity of SF reports, as with a larger geographical encompass, there presents more opportunity for flooding. Concerning YCOR and XCOR, the factors relate to the location of the zip code, according to its centroid. Specifically, the YCOR represents latitude (south to north directions), and the XCOR represents longitude (west to east direction). When viewing Fig 5h, the greater values indicate a northern direction, and the plot appears to indicate that there are lower complaints in the neighborhoods of northern NYC. The results strengthen this assertion, as the YCOR has high placement among predictors (10.38% importance in Model 1); thus, southern zip codes in NYC are shown to have greater street flooding susceptibility. Next, concerning XCOR, the higher values indicate a more eastward direction, and there is indication that there may be more street flooding complaints in the eastern sections of NYC; however, while a top 15 predictor, it is in the lower portion at 3.11% in Model 1. Thus, a west to east directionality is not as significant. While land feature and topographical conditions may be similar in zip codes of the same region, and thus, the effect of coordinates in SF reports may be due to these similarities, there may be additional reasons for geographical location showing effect. For instance, some locations are susceptible to sea level rise, a causal factor, known to increase flood risk (City of New York, 2022d). Hence, it may of interest to explore sea level rise in NYC for future studies. Lastly, PPDN has effect. By viewing Fig 5j, it appears that areas of greater population density have lower complaints. A hypothesis may be that, in NYC, more sophisticated drainage systems or systems with higher capacities are implemented in areas with a higher concentration of people; however, more investigation would be needed to substantiate the theory. Overall, the findings of the first analysis show that the physical and locational attributes, AREA and YCOR, account for considerable percent importance. When the simulations are run in Model 1, YCOR and AREA have 10.38% and 8.62%, respectively, percent importance. PPDN, on the other hand, remains in the lower portion of the top 15 predictors. Thus, it is seen that the size and location of a zip code have noticeable significance on SF reporting; in addition, PPDN has effect, but at a smaller extent.

5.3 Climatic Factors

Decidedly, precipitation is a preliminary cause of urban flooding and a driving force behind SF report filing. As this study is focusing on the spatial variability of SF reports, the evaluation of the effects of precipitation is dependent on precipitation pattern dynamics within NYC. In Model 1, neither of the seven rainfall variables present in the top 15 of predictors. It may be reasoned that rainfall differences within the NYC radar measurements are not of sufficient significance [in comparison to the other variables] to incur regional street flooding variations. Thus, precipitation spatial variability is seen to have a low effect on the spatial variability of SF reporting in NYC. Subsequently, the finding is a useful contribution toward modeling endeavors, as forecasted rainfall amounts may not suffice when identifying localized areas of increased flooding risk within NYC, as variability among zip codes appears to be due to other specific conditions within the neighborhood.

5.4 Socio-demographic Factors

Socio-demographics have appreciable influence towards the spatial variability of SF reports. Five of the 15 top predictors are socio-demographic for the Model 1. Specifically, COM02 (commuting to work - drove alone) comprises the largest percent importance at 17.79%. This variable signifies that the condition of driving to work has explanatory power towards SF reporting. An inference of this finding is that drivers are more endangered by street flooding and thus are more likely to file a complaint. Indeed, studies have shown that vehicular-related deaths comprise the majority of flooding fatalities in the United States (Ashley & Ashley, 2008; Han & Sharif, 2020). Furthermore, in a city, such as NYC, where certain regions have many subway stations, a large number of commuters are able to avoid vehicles; hence, they are not imminently confronted by this danger and do not report. Another possibility for the COM02 showing importance is that in suburban districts, where people drive more, there may be less sophisticated drainage systems or systems with lower capacities. In this type of instance, the COM02 factor may not directly motivate the variability of SF reports, as it may be a symptom of a different root cause. This inference is strengthened by the fact that many commuters who drive alone have a work location in a different borough or zip code. In fact, only 29% of those who work in Manhattan live in Manhattan, where 45% are residents of outer boroughs (City of New York, 2019). Thus, for instance, if a commuter from Staten Island drives to Manhattan and observes flooding in Manhattan, the flooding report would be that of the Manhattan location. Yet, the results show that commuters who drive are reporting flooding in their own zip codes. This gives further credence that the flooding is, indeed, taking place in their respective neighborhoods. Hence, in the case of the commuter who traverses zip codes, the bias may exist; however, it may not have a false skew, as the reported location is representative. Nevertheless, as there is significance with this variable, further research into commuter bias on crowdsourced flooding data may be beneficial, and an assignment of a weighting metric may be needed.

Overall, the socio-demographic variables within the top 15 comprise 38.94% of relative importance. In addition to COM02, homeownership variables, MORT02 (mortgage status - owner-occupied units - housing units with a mortgage) and HTEN02 (housing tenure - owner-occupied), show impact. Homeowners and homeowners with a mortgage, have a combined percent importance of 11.25%. The other socio-demographic characteristics include being employed in the armed service and living abroad the year prior. The connection of these variables warrants further investigation. Yet, it is seen that the crowdsourced data may be affected by the background and living characteristics of those who file, and when utilizing the data in urban flood research, further processing may be necessary.

5.5 The Influence of Catch Basins

Catch basins are a primary source for stormwater removal, and thus, blocked inlet drains contribute to street flooding, as rainwater, unable to infiltrate the impervious streets or enter the sewer system, may only ascend. While the mechanism of catch basin clogging is apparent, it is essential to assess whether a metropolitan has clogging to the extent of exasperation. Thus, to examine the effect of clogged catch basin issues in NYC, in Model 2, RF regressions are utilized, where CB reports are added as a predictor to the top 15 explanatory variables, and SF reports serve as the response. The results show that CB has 41.03% percent importance, where the second highest ranked predictor is at 6.73%; thus, in comparison, CB represents an overwhelming portion of significance in explaining the differences in SF reporting within zip codes. A clogged catch basin signifies a maintenance issue, where preventive actions, such as clearing the grates, or increasing public awareness, particularly in the advent of a rain event, would aid in remediation. Moreover, from 2010 to 2019 (the period of this study), the NYC DEP performed catch basin inspections every one to three years (DEP, 2020). Consequently, a plausible recommendation may to decrease the time between inspections to improve the issue of catch basin blockages. Therefore, the finding of CB as a strong predictor may provide direction for city management in flood relief, inspection scheduling and street cleaning measures.

Nonetheless, CB reports and SF reports do not hold a 1-to-1 linear relation, as depicted in Figure 6p. For example, if 750 CB reports are examined on the plot, the range of SF may be anywhere from 100 to almost 400 reports. The reason the relationship is not inherent is that SF and CB may occur independently. This separate occurrence is well illustrated by the September 1, 2021 urban flooding event from post tropical depression, Ida. The highest SF complaints, at 31 reports, had only two CB complaints; likewise, the zip code with the highest CB complaints of 10, had only 2 SF complaints. The disconnection between CB and SF occurred throughout the majority of zip codes for that day (The scatterplot depicting the SF and CB complaints for the day is in Appendix Fig D). Thus, a street flooding event is not always caused by a clogged catch basin, especially during high intensity rainfall days. Likewise, there may be water ponding near a clogged basin, where the streets are not flooded to an extent that instigates a SF report. Despite the nonlinearity, in NYC, the clogging of a catch basin does provide significant explanatory power for the street flooding variability among different neighborhoods.

5.6 The Effect of Zip Code Size

The results of the RF regressions show that the zip code size appears to have a strong effect. Of the predictors, it is seen that AREA, BLD, and FP are very significant in all the models of the study. In the simulations of only top predictors, AREA was ranked in the top five and found to have the relative importance of 8.62% in Model 1. In Model 2, once CB was added as a predictor, AREA had an importance of 5.12% and was amongst the top 10 predictors. Concerning the building factors, there is a relation with size, as the maps of total area of building footprints per zip code and total area of building footprints per square area per zip code show contrasting extents of saturation (See Fig 1e and Fig 1f). Thus, in the consideration that zip code area has influence, a complaint frequency analysis may be conducted. Per zip code, the SF complaints over the ten-year period is summed and then divided by the respective zip code area. This map is shown in Fig 1g. This frequency analysis, controlling for zip code size, visually pinpoints areas of high SF complaint density.

5.7 Model Limitations

Previous urban flood research has employed the RF algorithm (Chen et al., 2020; Feng et al., 2015; Kim & Kim, 2020; Lee et al., 2017; Sadler et al., 2018). Indeed, RF has been used explicitly in the evaluation of contributing factors for urban flooding. In Chen et al., RF methods were used in assessing explanatory variables, such as slope, land-use, rainfall, and altitude. However, the land-use category only distinguished between residential, water, grassland, farmland, and forest areas. Divergently, this study does not consolidate the feature class; yet, it seeks to understand the variations within. Thus, differences among the urban environment, such as building footprints and impervious cover, are explored. Moreover, this paper includes additional types of factors, such as catch basin clogging issues and population density. In another study, Sadler et al., RF is used to evaluate factor significance while also importing crowdsourced data. Similarly, this study applies citizen generated data; however, there is a greater number of variables incorporated. Sadler et al. includes environmental inputs, such as groundwater table level, tide, and wind; while this study considers topography, such as slope and elevation, in addition to infrastructural, land feature, and socio-demographical attributes.

Overall, this study is novel in its approach of using the RF machine learning technique, in conjunction with citizen collaborated data, in its evaluation of an encompassing and diverse dataset of predictors. As a measure of model skill, R^2 values are included. It is seen that when the number of predictors is minimized, R^2 values increase. For instance, when SF serves as the response, and there is the narrowing from 141 to 15 predictors, the R^2 increases to 0.63 from 0.58. Also noteworthy is that by adding CB to the top 15 predictors, with SF serving as the response, the R^2 increases to 0.71 from 0.63. Therefore, the inclusion of the infrastructural component compliments the explanatory power. The model may have been limited by the quantity of SF reports. This has been shown in Agonafir et al., where the results of negative binomial generalized linear regression model had higher R^2 values in zip codes with greater amounts of complaints. In addition, in the model where CB serves as the response (Model 3 of Appendix C), the R^2 is higher by 15 percentage points. This may have been due to a greater number of CB complaints being filed (85,607) as compared to SF complaints (25,574). Hence, as more crowdsourced data appears to reduce variability, increasing public awareness of the 311 platform may be a benefit to modeling endeavors.

5.8 Overall Synthesis

A summary of results is presented in Table 3. It is seen that when catch basin reports are added as a predictor towards street flooding reports (Model 2), they comprise nearly half the overall percent importance (41.13%). Moreover, the considerable contribution of the socio-demographic variables suggest that the crowdsourced data may be biased towards certain backgrounds. On the other hand, the relative importance of the land feature, topographical, and physical characteristics illuminates the specific factors affecting NYC street spatial variability. Thus, the results of this study aid in the identification of important variables in NYC street flooding, in addition to providing a directive for weighting assignments, which may be useful in urban risk zones mapping and prediction models.

The predictors appearing in Table 1 are the highest of their respective categories. Of the land feature and topographic category, SLPE, ELEV, GREEN, CBPA, IMPV, BLD, FP and FBPD are evaluated, and only GREEN and CBPA are not amongst the top predictors in either model. In Model 1, the top land feature and topographical elements aggregate to 35.14%. However, in Model 2, once the CB variable is added, the sum of these features is only 24.91 %. Of the physical characteristics and population dynamics category, XCOR, YCOR, AREA, PPDN, and POP are the included variables in the prescreening, and only POP is not ranked in the top 15 of predictors. In Model 1, the sum of the percent importance of the top physical characteristics and population dynamics factors is 26.15%.

Again, once CB is added in Model 2, the total importance of the variables decreases to 14.62%. Regarding the climatic category, seven variables are input for the prescreening analysis, NZMN, NZKT, NZSW, NZSD, PERC, MNWTS, and MXWTS; yet, for Model 1, and by consequence, Model 2, none of the precipitation parameters ranked in the top 15 of predictors. Lastly, when viewing the total listing socio-demographic variables (including those in the initial prescreen), it is seen that seven of the 121 variables appear as top predictors in either model. These variables include COM02, EMP06, RES04, MORT02, LANG02, HTEN02 and INC10. In Model 1, the variables total to 38.94%; then in Model 2, the total reduces to 19.44%, with the addition of the CB predictor. Thus, the RF models successfully signified factors from each of the input types within the study, of which influence the spatial variability of SF and CB reports within NYC zip codes.

Key Predictors			
Land Feature and Topographical		**Model 1**	**Model 2**
BLD	Number of buildings	**9.43%**	**3.47%**
FP	Sum of the building footprints	**6.10%**	**3.43%**
SLPE	Slope - mean percent rise	**7.61%**	**6.73%**
ELEV	Mean elevation	**5.10%**	**2.12%**
IMPV	Percent of impervious Cover	**3.68%**	**4.37%**
FPBD	Sum of the building footprints per unit area	**3.22%**	**4.79%**
Physical Characteristics and Population Dynamics			
AREA	Area		

		8.62%	**5.12%**
YCOR	Centroid of y coordinate	**10.38%**	**6.19%**
XCOR	Centroid of x coordinate	**3.11%**	**0.02%**
PPDN	Population Density	**4.04%**	**3.29%**
Socio-demographic			
COM02	COMMUTING TO WORK - Car, truck, or van -- drove alone	**17.79%**	**6.56%**
EMP06	EMPLOYMENT STATUS - In labor force - Armed Forces	**6.58%**	**4.53%**
RES04	RESIDENCE 1 YEAR AGO - Population 1 year and over - Abroad	**3.32%**	**3.44%**
MORT02	MORTGAGE STATUS - Owner-occupied units - Housing units with a mortgage	**6.12%**	**3.40%**
HTEN02	HOUSING TENURE -	**5.13%**	**1.52%**

	Occupied housing units - Owner-occupied		
Infrastructural			
CB	Total catch basin complaints		**41.13%**

Table 7 A summary showing the complete listing of the top 15 predictors of all models. The percent importance values for Model 1 are the median values of 50 simulation runs for the top 15 predictors only. The percent importance values for Model 2 are the median values of 50 simulation runs for the 16 predictors.

6 Conclusions

Urban flood research is presented with the complexities of the urban environment. The physical and social characteristics of a sprawling metropolitan are oftentimes dynamic - varying from one neighborhood to the next. This is especially evident in NYC, where diversity is prevalent. The land features range from high-rise buildings in impervious areas to residential neighborhoods with parks and ponds; the topography fluctuates from hilly and steep in some places to flat and low-lying in others; and, the people of NYC vary in background, income, and commuting style. As there are multiple factors influencing the behavior of runoff, a distinct feature of a neighborhood may have contribution, and thus, there is a need for analysis. Subsequently, a model with the ability to accommodate these intricacies is of value.

This paper implements the Random Forest machine learning algorithm to evaluate the spatial variability of NYC crowdsourced street flooding reports. A chief benefit of the model is the incorporation of a large dataset of land feature, topographical, physical and population, socio-demographic, locational and climatic variables to produce an output of predictor importance for each variable. The results of this study show that land feature characteristics, such as the number of buildings and building footprint area, affect the differences in street flood reporting per zip code. In addition, slope is a signified factor, and the location and the size of the zip code also influenced the frequency of street flood reporting. Furthermore, a major finding is that catch basin clogged reports, once added as a predictor, has the highest relative importance. As such, improved street cleaning methods or increased inspections may be recommended. Moreover, this study is the first of its kind to evaluate the role of socio-demographics towards NYC 311 street flooding and catch basin reporting behavior. With this analysis, it is found that the 311 street flooding data appears to be skewed by commuters who drive to work, rather than those who use alternative modes of transportation. Thus, methods of filtering bias may be needed when importing the citizen generated data in urban flood modeling. Overall, this paper presents the factors significant in the regional variations of NYC street flood reporting.

7 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8 Data Availability

The sources of the data (311 complaints) are available here: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.

Radar data may be accessed here: https://data.eol.ucar.edu/cgi-bin/codiac/fgr_form/id=21.093.

The processed data and the codes used in this study are available from the corresponding authors upon reasonable request.

9 Acknowledgement

This research was supported by NOAA-CESSRST Cooperative Agreement (NOAA/EPP Grant # NA16SEC4810008). The statements contained within the manuscript are not the opinions of the funding agency or the U.S. government but reflect the authors' opinions.

10 References

Agonafir, C., Ramirez Pabon, A., Lakhankar, T., Khanbilvardi, R., & Devineni, N. (2021). Understanding New York City Street Flooding through 311 Complaints. *Journal of Hydrology*, *605*(March 2021), 127300. https://doi.org/10.1016/j.jhydrol.2021.127300

Albers, S. J., Dery, S. J., & Petticrew, E. L. (2015). Flooding in the Nechako River Basin of Canada: A random forest modeling approach to flood analysis in a regulated reservoir system. *Canadian Water Resources Journal*, *41*.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278.

AlThuwaynee, O. F., Kim, S.-W., Najemaden, M. A., Aydda, A., Balogun, A.-L., Fayyadh, M. M., & Park, H.-J. (2021). Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. *Environmental Science and Pollution Research*, *28*(32), 43544–43566. https://doi.org/10.1007/s11356-021-13255-4

Aon. (2021). Weather, Climate and Catastrophe Insight. https://www.aon.com/getmedia/1b516e4d-c5fa-4086-9393-5e6afb0eeded/20220125-2021-weather-climate-catastrophe-insight.pdf.aspx

Ashley, S. T., & Ashley, W. S. (2008). Flood fatalities in the United States. *Journal of Applied Meteorology and Climatology*, *47*(3), 805–818. https://doi.org/10.1175/2007JAMC1611.1

Asadieh, B., & Krakauer, N. (2016). Impacts of Changes in Precipitation Amount and Distribution on Water Resources Studied Using a Model Rainwater Harvesting System. *Journal of the American Water Resources Association*.

Bado, V. B., & Bationo, A. (2018). Integrated Management of Soil Fertility and Land Resources in Sub-Saharan Africa: Involving Local Communities. *Advances in Agronomy*, *150*, 1–33. https://doi.org/10.1016/BS.AGRON.2018.02.001

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, *18*(3), 257–279. https://doi.org/10.1007/s10588-012-9121-2

Basiri, A., Haklay, M., Foody, G., & Mooney, P. (2019). Crowdsourced geospatial data quality: challenges and future directions. *International Journal of Geographical Information Science*, *33*(8), 1588–1593. https://doi.org/10.1080/14158816.2019.1593422

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*.

Bruwier, M., Maravat, C., Mustafa, A., Teller, J., Pirotton, M., Erpicum, S., Archambeau, P., & Dewals, B. (2020). Influence of urban forms on surface flow in urban pluvial flooding. *Journal of Hydrology*, *582*(December 2019). https://doi.org/10.1016/j.jhydrol.2019.124493

Bureau of Economic Analysis. (2021). *Metropolitan Statistical Areas, GDP, 2020*. https://apps.bea.gov/iTable/iTable.cfm?reqid=99&step=1#reqid=99&step=1

Bulti, D. T., & Abebe, B. G. (2020). A review of flood modeling methods for urban pluvial flood application. In *Modeling Earth Systems and Environment* (Vol. 6, Issue 3, pp. 1293–1302). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s40808-020-00803-z

Chang, T. J., Wang, C. H., & Chen, A. S. (2015). A novel approach to model dynamic flow interactions between storm sewer system and overland surface for different land covers in urban areas. *Journal of Hydrology*, *524*.

Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Wang, X., Bian, H., Zhang, S., Pradhan, B., & Ahmad, B. bin. (2020). Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods. *Science of The Total Environment*, *701*, 134979. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.134979

City of New York. (2019). *The Ins and Outs of NYC Commuting*. https://www1.nyc.gov/assets/planning/download/pdf/planning-level/housing-economy/nyc-ins-and-out-of-commuting.pdf

City of New York. (2022a). *About NYC 311*. https://portal.311.nyc.gov/about-nyc-311/

City of New York. (2022b). *Catch Basin Complaint*. https://portal.311.nyc.gov/article/?kanumber=KA-01084

City of New York. (2022c). *Flood Prevention*. https://www1.nyc.gov/site/dep/environment/flood-prevention.page

City of New York. (2022d). *Info Brief: Flood Risk in NYC*. https://www1.nyc.gov/assets/planning/download/pdf/plans-studies/climate-resiliency/flood-risk-nyc-info-brief.pdf

City of New York. (2022e). *Street Flooding*. https://portal.311.nyc.gov/article/?kanumber=KA-02198

City of New York. (2022f). *Green Roofs & Solar Panels*. https://www1.nyc.gov/site/buildings/property-or-business-owner/green-roofs-solar-panels.page

Comber, A., Mooney, P., Purves, R. S., Rocchini, D., & Walz, A. (2016). Crowdsourcing: It Matters Who the Crowd Are. The Impacts of between Group Variations in Recording Land Cover. *PLOS ONE*, *11*(7), e0158329. https://doi.org/10.1371/journal.pone.0158329

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., & Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. *Journal of Hydrology*, *569*.

Dede, M., Widiawaty, M. A., Pramulatsih, G. P., Ismail, A., Ati, A., & Murtianto, A. (2019). Integration of Participatory Mapping, Crowdsourcing and Geographic Information System in Flood Disaster Management (Case Study Ciledug Lor, Cirebon). *Journal of Information Technology and Its Utilization*, *2*.

Diakakis, M., Deligiannakis, G., Pallikarakis, A., & Skordoulis, M. (2016). Factors controlling the spatial distribution of flash flooding in the complex environment of a metropolitan urban area. The case of Athens 2013 flash flood event. *International Journal of Disaster Risk Reduction*, *18*, 171–180. https://doi.org/https://doi.org/10.1016/j.ijdrr.2016.06.010

Dietz, M. E. (2007). Low Impact Development Practices: A Review of Current Research and Recommendations for Future Directions. *Water, Air, and Soil Pollution*.

Dixon, B., Johns, RebeccaA., & Fernandez, A. (2021). The role of crowdsourced data, participatory decision-making and mapping of flood related events. *Applied Geography*, *128*, 102393. https://doi.org/https://doi.org/10.1016/j.apgeog.2021.102393

Dou, X., Song, J., Wang, L., Tang, B., Xu, S., Kong, F., & Jiang, X. (2018). Flood risk assessment and mapping based on a modified multi-parameter flood hazard index model in the Guanzhong Urban Area, China. *Stochastic Environmental Research and Risk Assessment*, *32*(4), 1131–1146. https://doi.org/10.1007/s00477-017-1429-5

Du, J. (2011). *NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data. Version 1.0. UCAR/NCAR - Earth Observing Laboratory*. https://doi.org/10.5065/D6PG1QDD

Feng, Q., Liu, J., & Gong, J. (2015). Urban Flood Mapping Based on Unmanned Aerial Vehicle Remote Sensing and Random Forest Classifier—A Case of Yuyao, China. In *Water* (Vol. 7, Issue 4). https://doi.org/10.3390/w7041437

Hamidi, A., Devineni, N., Booth, J. F., Hosten, A., Ferraro, R. R., & Khanbilvardi, R. (2017). Classifying urban rainfall extremes using weather radar data: An application to the greater New York Area. *Journal of Hydrometeorology*, *18*(3), 611–623. https://doi.org/10.1175/JHM-D-16-0193.1

Han, Z., & Sharif, H. O. (2020). Vehicle-Related Flood Fatalities in Texas, 1959–2019. In *Water* (Vol. 12, Issue 10). https://doi.org/10.3390/w12102884

Hanchey, A., Schnall, A., Bayleyegn, T., Jiva, S., Khan, A., Siegel, V., Funk, R., & Svendsen, E. (2021). *Notes from the Field: Deaths Related to Hurricane Ida Reported by Media — Nine States, August 29–September 9, 2021*. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/70/wr/mm7039a3.htm

Hedges, M., & Dunn, S. (2018). Crowdsourcing and memory. *Academic Crowdsourcing in the Humanities*, 127–145. https://doi.org/10.1016/B978-0-08-100941-3.00008-5

Helmrich, A. M., Ruddell, B. L., Bessem, K., Chester, M. v., Chohan, N., Doerry, E., Eppinger, J., Garcia, M., Goodall, J. L., Lowry, C., & Zahura, F. T. (2021). Opportunities for crowdsourcing in urban flood monitoring. *Environmental Modelling & Software*, *143*, 105124. https://doi.org/10.1016/J.ENVSOFT.2021.105124

Huang, J. C., Tsai, Y. C., Wu, P. Y., Lien, Y. H., Chien, C. Y., Kuo, C. F., Hung, J. F., Chen, S. C., & Kuo, C. H. (2020). Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. *Computer Methods and Programs in Biomedicine*, *195*, 105536. https://doi.org/10.1016/J.CMPB.2020.105536

Impact of NYW Bonds. (n.d.). https://www1.nyc.gov/site/nyw/investing-in-nyw-bonds/the-impact-of-investing.page

Jato-Espino, D., Lobo, A., & Ascorbe-Salcedo, A. (2019). Urban flood risk mapping using an optimised additive weighting methodology based on open data. *Journal of Flood Risk Management*, *12*(S1).

Kandilioti, G., & Makropoulos, C. (2012). Preliminary flood risk assessment: the case of Athens. *Natural Hazards*, *61*(2), 441–468. https://doi.org/10.1007/s11069-011-9930-5

Kelleher, C., & McPhillips, L. (2020). Exploring the application of topographic indices in urban areas as indicators of pluvial flooding locations. *Hydrological Processes*. https://doi.org/10.1002/hyp.14128

Kim, H. il, & Kim, B. H. (2020). Flood Hazard Rating Prediction for Urban Areas Using Random Forest and LSTM. *KSCE Journal of Civil Engineering*, *24*(12), 3884–3896. https://doi.org/10.1007/s12205-020-0951-z

Leandro, J., Schumann, A., & Pfister, A. (2016). A step towards considering the spatial heterogeneity of urban key features in urban hydrology flood modelling. *Journal of Hydrology*, *535*.

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, *8*(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lin, J., He, X., Lu, S., Liu, D., & He, P. (2021). Investigating the influence of three-dimensional building configuration on urban pluvial flooding using random forest algorithm. *Environmental Research*, *196*, 110438. https://doi.org/10.1016/J.ENVRES.2020.110438

Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In B. Liu, M. Ma, & J. Chang (Eds.), *Information Computing and Applications*. Springer.

Loh, W. (2004). Regression Trees with. Unbiased Variable Selection. *Korean Journal of Applied Statistics*, *17*(3), 459–473. https://doi.org/10.5351/kjas.2004.17.3.459

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

MathWorks. (2022). *Select Predictors for Random Forest*. https://www.mathworks.com/help/stats/select-predictors-for-random-forests.html

Minkoff, S. L. (2015). NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City. *Urban Affairs Review*.

Moreno, P. G., Artes-Rodríguez, A., Teh, Y. W., & Perez-Cruz, F. (2015). Bayesian nonparametric crowdsourcing. *Journal of Machine Learning Research*, *16*, 1607–1627.

National Academies of Sciences, Engineering, and M. (2019). *Framing the Challenge of Urban Flooding in the United States*. https://doi.org/https://doi.org/10.17226/25381

Novikov, S. L. (1981). Elevation: A major influence on the hydrology of new hampshire and vermont, usa. *Hydrological Sciences Bulletin*, *26*(4), 399–413. https://doi.org/10.1080/02626668109490904

NWS. (2022a). *Flash Flood Guidance*. https://www.weather.gov/mbrfc/ffg_alt

NWS. (2022b). *National Weather Service New York, NY Watch Warning Advisory Definitions Page*. https://www.weather.gov/okx/wwa_definitions

Ouma, Y. O., & Tateishi, R. (2014a). Urban Flood Vulnerability and Risk Mapping Using Integrated Multi-Parametric AHP and GIS: Methodological Overview and Case Study Assessment. In *Water* (Vol. 6, Issue 6). https://doi.org/10.3390/w6061515

Pak, B., Chua, A., & vande Moere, A. (2017). FixMyStreet Brussels: Socio-Demographic Inequality in Crowdsourced Civic Participation. *Journal of Urban Technology*, *24*(2), 65–87. https://doi.org/10.1080/10630732.2016.1270047

Plumer, B. (2021). Flooding From Ida Kills Dozens of People in Four States. *The New York Times*. https://www.nytimes.com/live/2021/09/02/nyregion/nyc-storm

Rahmati, O., Darabi, H., Panahi, M., Kalantari, Z., Naghibi, S. A., Ferreira, C. S. S., Kornejady, A., Karimidastenaei, Z., Mohammadi, F., Stefanidis, S., Tien Bui, D., & Haghighi, A. T. (2020). Development of novel hybridized models for urban flood susceptibility mapping. *Scientific Reports*, *10*(1), 1–19. https://doi.org/10.1038/s41598-020-69703-7

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(1), 93–104. https://doi.org/10.1016/J.ISPRSJPRS.2011.11.002

Rusdah, D. A., & Murfi, H. (2020). XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*(8), 1336. https://doi.org/10.1007/s42452-020-3128-y

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

Scheuer, S., Haase, D., & Meyer, V. (2011). Exploring multicriteria flood vulnerability by integrating economic, social and ecological dimensions of flood risk and coping capacity: from a starting point view towards an end point view of vulnerability. *Natural Hazards*, *58*(2), 731–751. https://doi.org/10.1007/s11069-010-9666-7

See, L. (2019). A review of citizen science and crowdsourcing in applications of pluvial flooding. *Frontiers in Earth Science*, *7*(March), 1–7. https://doi.org/10.3389/feart.2019.00044

Serrano, S. E. (2010). *Hydrology for engineers, geologists, and environmental professionals: an integrated treatment of surface, subsurface, and contaminant hydrology*. Hydroscience Inc.

Sharif, H. O., Yates, D., Roberts, R., & Mueller, C. (2006). The use of an automated nowcasting system to forecast flash floods in an urban watershed. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM482.1

Smith, B., & Rodriguez, S. (2017). Spatial analysis of high-resolution radar rainfall and citizen-reported flash flood data in ultra-urban New York City. *Water (Switzerland)*. https://doi.org/10.3390/w9100736

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

S.V., C., Harindranathan, N., A, A., & N.S., A. (2015). Impacts of Impervious Surfaces on the Environment. *International Journal of Engineering Science Invention*, *4*(5), 27–31. www.ijesi.org

Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *CATENA*, *175*, 174–192. https://doi.org/https://doi.org/10.1016/j.catena.2018.12.011

Thorndahl, S., Einfalt, T., Willems, P., Ellerbæk Nielsen, J., ten Veldhuis, M. C., Arnbjerg-Nielsen, K., Rasmussen, M. R., & Molnar, P. (2017). Weather radar rainfall data in urban hydrology. *Hydrology and Earth System Sciences*, *21*(3), 1359–1380. https://doi.org/10.5194/hess-21-1359-2017

United States Census Bureau. (2012). *Largest Urbanized Areas With Selected Cities and Metro Areas*. https://www.census.gov/dataviz/visualizations/026/

Viessman, Lewis, G. L. (2003). *Introduction to Hydrology* (Fifth). Pearson Education, Inc.

Wang, R. Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, *111*, 139–147. https://doi.org/10.1016/J.CAGEO.2017.11.008

Wang, X., Kingsland, G., Poudel, D., & Fenech, A. (2019a). Urban flood prediction under heavy precipitation. *Journal of Hydrology*, *577*.

Wang, X., Kingsland, G., Poudel, D., & Fenech, A. (2019b). Urban flood prediction under heavy precipitation. *Journal of Hydrology*, *577*, 123984. https://doi.org/10.1016/J.JHYDROL.2019.123984

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

Xu, Z., & Wang, Z. (2019). A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)*, 278–283. https://doi.org/10.1109/ICACI.2019.8778622

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016). Simulating California reservoir operation using theclassification and regression-tree algorithm combinedwith a shuffled cross-validation scheme. *Water Resources Research*.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, *16*(3), 417–434. https://doi.org/10.1007/s10796-012-9350-4

11 Appendix

11.1 Appendix A

Table A.1 A list of the Socio-demographic variables and respective abbreviations

Abbreviation	Socio-demographic Variable
CITZ01	U.S. CITIZENSHIP STATUS - Foreign-born population
CITZ02	U.S. CITIZENSHIP STATUS - Foreign-born population - Naturalized U.S. citizen
CITZ03	U.S. CITIZENSHIP STATUS - Foreign-born population - Not a U.S. citizen
COM01	COMMUTING TO WORK - Workers 16 years and over

COM02	COMMUTING TO WORK - Workers 16 years and over - Car, truck, or van -- drove alone
COM03	COMMUTING TO WORK - Workers 16 years and over - Car, truck, or van -- carpooled
COM04	COMMUTING TO WORK - Workers 16 years and over - Public transportation (excluding taxicab)
COM05	COMMUTING TO WORK - Workers 16 years and over - Walked
COM06	COMMUTING TO WORK - Workers 16 years and over - Other means
COM07	COMMUTING TO WORK - Workers 16 years and over - Worked at home
COM08	COMMUTING TO WORK - Workers 16 years and over - Mean travel time to work (minutes)
EDU01	EDUCATIONAL ATTAINMENT - Population 25 years and over
EDU02	EDUCATIONAL ATTAINMENT - Population 25 years and over - Less than 9th grade
EDU03	EDUCATIONAL ATTAINMENT - Population 25 years and over - 9th to 12th grade, no diploma
EDU04	EDUCATIONAL ATTAINMENT - Population 25 years and over - High school graduate (includes equivalency)
EDU05	EDUCATIONAL ATTAINMENT - Population 25 years and over - Some college, no degree
EDU06	EDUCATIONAL ATTAINMENT - Population 25 years and over - Associate's degree
EDU07	EDUCATIONAL ATTAINMENT - Population 25 years and over - Bachelor's degree
EDU08	EDUCATIONAL ATTAINMENT - Population 25 years and over - Graduate or professional degree
EDU09	EDUCATIONAL ATTAINMENT - Population 25 years and over - High school graduate or higher
EDU10	EDUCATIONAL ATTAINMENT - Population 25 years and over - Bachelor's degree or higher
EMP01	EMPLOYMENT STATUS - Population 16 years and over
EMP02	EMPLOYMENT STATUS - Population 16 years and over - In labor force
EMP03	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Civilian labor force
EMP04	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Civilian labor force – Employed
EMP05	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Civilian labor force -

	Unemployed
EMP06	EMPLOYMENT STATUS - Population 16 years and over - In labor force - Armed Forces
EMP07	EMPLOYMENT STATUS - Population 16 years and over - Not in labor force
EMP08	EMPLOYMENT STATUS - Civilian labor force
FERT01	FERTILITY - Number of women 15 to 50 years old who had a birth in the past 12 months
GEOID2	Zip Code ID
HISL01	HISPANIC OR LATINO AND RACE - Total population
HISL02	HISPANIC OR LATINO AND RACE - Total population - Hispanic or Latino (of any race)
HISL03	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino
HISL04	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - White alone
HISL05	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - Black or African American alone
HISL06	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - American Indian and Alaska Native alone
HISL07	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - Asian alone
HISL08	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - Native Hawaiian and Other Pacific Islander alone
HISL09	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - Some other race alone
HISL10	HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino - Two or more races
HOC01	HOUSING OCCUPANCY - Total housing units
HOC02	HOUSING OCCUPANCY - Total housing units - Occupied housing units
HOC03	HOUSING OCCUPANCY - Total housing units - Vacant housing units
HOC04	HOUSING OCCUPANCY - Total housing units - Homeowner vacancy rate
HOC05	HOUSING OCCUPANCY - Total housing units - Rental vacancy rate
HSHD01	HOUSEHOLDS BY TYPE - Total households

HSHD02	HOUSEHOLDS BY TYPE - Total households - Family households (families)
HSHD03	HOUSEHOLDS BY TYPE - Total households - Family households (families) - With own children of the householder under 18 years
HSHD04	HOUSEHOLDS BY TYPE - Total households - Average household size
HSHD05	HOUSEHOLDS BY TYPE - Total households - Average family size
HTEN01	HOUSING TENURE - Occupied housing units
HTEN02	HOUSING TENURE - Occupied housing units - Owner-occupied
HTEN03	HOUSING TENURE - Occupied housing units - Renter-occupied
HVAL01	VALUE - Owner-occupied units - Median (dollars)
INC01	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households
INC02	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - Less than $10,000
INC03	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $10,000 to $14,999
INC04	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $15,000 to $24,999
INC05	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $25,000 to $34,999
INC06	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $35,000 to $49,999
INC07	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $50,000 to $74,999
INC08	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $75,000 to $99,999
INC09	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $100,000 to $149,999
INC10	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households

	- $150,000 to $199,999
INC11	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $200,000 or more
INC12	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - Median household income (dollars)
INC13	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - Mean household income (dollars)
LANG01	LANGUAGE SPOKEN AT HOME - Population 5 years and over
LANG02	LANGUAGE SPOKEN AT HOME - Population 5 years and over - English only
LANG03	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Language other than English
MORT01	MORTGAGE STATUS - Owner-occupied units
MORT02	MORTGAGE STATUS - Owner-occupied units - Housing units with a mortgage
MORT03	MORTGAGE STATUS - Owner-occupied units - Housing units without a mortgage
RENT01	GROSS RENT - Occupied units paying rent
RENT02	GROSS RENT - Occupied units paying rent - Less than $500
RENT03	GROSS RENT - Occupied units paying rent - $500 to $999
RENT04	GROSS RENT - Occupied units paying rent - $1,000 to $1,499
RENT05	GROSS RENT - Occupied units paying rent - $1,500 to $1,999
RENT06	GROSS RENT - Occupied units paying rent - $2,000 to $2,499
RENT07	GROSS RENT - Occupied units paying rent - $2,500 to $2,999
RENT08	GROSS RENT - Occupied units paying rent - $3,000 or more
RENT09	GROSS RENT - Occupied units paying rent - Median (dollars)
RENT10	GROSS RENT - Occupied units paying rent - No rent paid
RES01	RESIDENCE 1 YEAR AGO - Population 1 year and over
RES02	RESIDENCE 1 YEAR AGO - Population 1 year and over - Same house
RES03	RESIDENCE 1 YEAR AGO - Population 1 year and over - Different house in the U.S.
RES04	RESIDENCE 1 YEAR AGO - Population 1 year and over - Abroad

SCH01	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school
SCH02	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - Nursery school, preschool
SCH03	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - Kindergarten
SCH04	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - Elementary school (grades 1-8)
SCH05	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - High school (grades 9-12)
SCH06	SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - College or graduate school
SXAG01	SEX AND AGE - Total population
SXAG02	SEX AND AGE - Total population - Male
SXAG03	SEX AND AGE - Total population - Female
SXAG04	SEX AND AGE - Total population - Under 5 years
SXAG05	SEX AND AGE - Total population - 5 to 9 years
SXAG06	SEX AND AGE - Total population - 10 to 14 years
SXAG07	SEX AND AGE - Total population - 15 to 19 years
SXAG08	SEX AND AGE - Total population - 20 to 24 years
SXAG09	SEX AND AGE - Total population - 25 to 34 years
SXAG10	SEX AND AGE - Total population - 35 to 44 years
SXAG11	SEX AND AGE - Total population - 45 to 54 years
SXAG12	SEX AND AGE - Total population - 55 to 59 years
SXAG13	SEX AND AGE - Total population - 60 to 64 years
SXAG14	SEX AND AGE - Total population - 65 to 74 years
SXAG15	SEX AND AGE - Total population - 75 to 84 years
SXAG16	SEX AND AGE - Total population - 85 years and over
SXAG17	SEX AND AGE - Total population - Median age (years)

VOTE01	CITIZEN, VOTING AGE POPULATION - Citizen, 18 and over population
RACE01	RACE - Total population
RACE02	RACE - Total population - One race
RACE03	RACE - Total population - One race – White
RACE04	RACE - Total population - One race - Black or African American
RACE05	RACE - Total population - One race - American Indian and Alaska Native
RACE06	RACE - Total population - One race – Asian
RACE07	RACE - Total population - One race - Native Hawaiian and Other Pacific Islander
RACE08	RACE - Total population - One race - Some other race
RACE09	RACE - Total population - Two or more races

11.2 Appendix B

Specifically, as provided by the Mathworks documentation (found at https://www.mathworks.com/help/stats/regressionbaggedensemble.oobpermutedpredictorimportance.html#bvf92si-1), the model operates as follows:

- For each tree, t, of the prediction trees:

$$t = 1, \dots, 500$$

- Splitting the indices of the predictor variables to grow t and identifying OOB:

$$s_t \in \{1, \dots, p\}$$

where, p is the number of explanatory variables.

- The OOB error is estimated, e_t.
- For each explanatory variable $x_j, j \in s_t$:
  1. Observations of x_j are randomly permuted.
  2. By the OOB containing the permuted values of x_j, model error, e_{tj} is estimated.
  3. The difference is taken: $d_{tj} = e_{tj} - e_t$

- By the differences over the learners, $j = 1, \ldots, p$, the mean, $\overline{d_j}$, and standard deviation, σ_j for each explanatory variable are determined.
- The impOOB for x_j is calculated as $\frac{\overline{d_j}}{\sigma_j}$.

11.3 Appendix C

Table C.1 provide the median relative importance ratio for each of the 141 variables in the Model 1 RF simulations, where SF serves as the response. **Table C.2** provides the median relative importance ratio for each of the 141 variables in the Model 3 RF simulations, where CB serves as the response.

C.1 Results of the RF simulations with SF as the response		C.2 Results of the RF simulations with CB as the response	
Variables	Importance Ratio	Variables	Importance Ratio
COM02	0.0791	BLD	0.1243
BLD	0.0442	COM02	0.0499
YCOR	0.0397	FP	0.0419
AREA	0.0388	AREA	0.0286
FP	0.0320	MORT02	0.0268
MORT02	0.0257	HTEN02	0.0256
SLPE	0.0244	ELEV	0.0177
EMP06	0.0219	LANG02	0.0174
HTEN02	0.0206	YCOR	0.0167
ELEV	0.0150	PPDN	0.0147
RES04	0.0140	NZKT	0.0140
PPDN	0.0129	NZSW	0.0138
IMPV	0.0116	SLPE	0.0133
FPBD	0.0110	NZMN	0.0132
XCOR	0.0107	INC10	0.0119
HSHD01	0.0105	XCOR	0.0117
LANG02	0.0104	FPBD	0.0111
NZKT	0.0098	HISL04	0.0090
RACE08	0.0098	EDU05	0.0088
COM08	0.0098	INC11	0.0084
EDU05	0.0092	HSHD01	0.0081
HISL03	0.0092	RACE06	0.0080
SXAG12	0.0090	COM03	0.0080
NZMN	0.0085	HISL07	0.0079
LANG03	0.0085	RACE05	0.0078
EDU06	0.0085	COM05	0.0078
HISL02	0.0084	EDU04	0.0077
RENT05	0.0083	RACE03	0.0076
RACE09	0.0083	HISL09	0.0074
COM03	0.0079	RENT05	0.0073
EDU02	0.0079	RENT06	0.0072
NZSW	0.0076	RES04	0.0072
HTEN03	0.0074	MNWTS	0.0069
HISL09	0.0074	LANG03	0.0068
CITZ02	0.0074	EMP06	0.0068
SXAG06	0.0073	HSHD02	0.0067
RACE06	0.0073	IMPV	0.0066

HISL04	0.0073	HSHD04	0.0065
EDU04	0.0072	CBPA	0.0064
COM04	0.0072	EDU09	0.0064
SXAG13	0.0072	HISL06	0.0063
SXAG14	0.0069	INC12	0.0062
EDU07	0.0069	INC07	0.0062
CITZ03	0.0068	HSHD03	0.0061
HSHD02	0.0068	COM08	0.0060
RENT04	0.0064	SXAG14	0.0060
SCH04	0.0064	SXAG13	0.0060
RENT06	0.0063	SXAG12	0.0060
SCH05	0.0063	RES03	0.0059
RES03	0.0062	HISL03	0.0059
INC07	0.0061	PERC	0.0059
SXAG09	0.0061	INC05	0.0058
GREEN	0.0061	HVAL01	0.0057
CITZ01	0.0061	RENT04	0.0057
RACE05	0.0059	CITZ03	0.0056
HISL07	0.0059	RACE08	0.0055
INC05	0.0058	INC13	0.0055
PERC	0.0057	RENT07	0.0055
EDU03	0.0057	EDU07	0.0054
COM05	0.0056	HTEN03	0.0054
INC03	0.0056	CITZ01	0.0053
INC04	0.0056	INC03	0.0053
RACE03	0.0055	COM04	0.0052
SXAG02	0.0055	RENT08	0.0052
INC11	0.0055	INC04	0.0052
SXAG11	0.0054	INC09	0.0052
COM06	0.0053	HSHD05	0.0051
MNWTS	0.0053	SCH01	0.0050
SXAG05	0.0053	INC02	0.0049
INC10	0.0052	GREEN	0.0049
EDU09	0.0051	SXAG09	0.0048
HISL06	0.0050	SXAG05	0.0048
SXAG15	0.0050	EMP07	0.0047
POP	0.0049	EDU02	0.0046
EMP07	0.0048	COM06	0.0046
COM07	0.0048	HOC01	0.0046
SXAG10	0.0047	MORT03	0.0046
SXAG16	0.0047	VOTE01	0.0046
EMP05	0.0047	FERT01	0.0044
MORT03	0.0047	EDU01	0.0043
INC12	0.0046	SXAG10	0.0043
HOC01	0.0046	SCH04	0.0042
EMP02	0.0045	SCH05	0.0042
EDU10	0.0045	SXAG07	0.0042
INC08	0.0045	SXAG06	0.0041
RES02	0.0044	SXAG04	0.0041
EMP04	0.0043	CITZ02	0.0041
VOTE01	0.0043	RACE09	0.0041
HVAL01	0.0043	RES01	0.0041
SXAG03	0.0043	SXAG08	0.0040

RENT01	0.0042	SCH02	0.0040
INC09	0.0041	EDU10	0.0040
SCH02	0.0041	SXAG16	0.0039
HSHD03	0.0040	NZSD	0.0039
RACE02	0.0040	POP	0.0038
EMP01	0.0040	RACE02	0.0038
SXAG07	0.0039	COM01	0.0037
SXAG04	0.0039	RENT01	0.0037
EDU01	0.0039	INC08	0.0037
INC06	0.0039	SXAG02	0.0036
RENT08	0.0038	SXAG11	0.0036
HISL10	0.0038	MORT01	0.0036
SCH01	0.0038	SCH03	0.0034
HISL05	0.0037	EMP02	0.0034
SCH06	0.0037	SXAG03	0.0034
COM01	0.0035	SXAG01	0.0033
RENT10	0.0033	LANG01	0.0032
CBPA	0.0033	EDU03	0.0032
SXAG08	0.0031	RENT10	0.0031
RENT03	0.0031	EMP05	0.0030
LANG01	0.0031	EDU08	0.0030
MXWTS	0.0028	EMP01	0.0030
SCH03	0.0028	HOC05	0.0029
RES01	0.0027	RENT03	0.0029
RENT07	0.0027	RACE04	0.0029
HSHD04	0.0026	RES02	0.0029
HOC03	0.0025	EDU06	0.0028
RENT02	0.0023	EMP04	0.0027
NZSD	0.0019	INC01	0.0027
RACE04	0.0017	INC06	0.0025
INC02	0.0013	SXAG17	0.0025
INC13	0.0012	SXAG15	0.0024
SXAG17	0.0009	COM07	0.0024
FERT01	0.0009	MXWTS	0.0024
INC01	0.0007	HOC03	0.0024
RENT09	0.0005	SCH06	0.0020
HOC05	0.0003	RENT02	0.0018
HSHD05	0.0003	HISL10	0.0018
EDU08	0.0001	HISL02	0.0016
EMP03	0.0000	RENT09	0.0005
EMP08	0.0000	HISL05	0.0002
EMP09	0.0000	EMP03	0.0000
HOC02	0.0000	EMP08	0.0000
HOC04	0.0000	EMP09	0.0000
HTEN01	0.0000	HOC02	0.0000
MORT01	0.0000	HOC04	0.0000
SXAG01	0.0000	HTEN01	0.0000
RACE01	0.0000	RACE01	0.0000
RACE07	0.0000	RACE07	0.0000
HISL01	0.0000	HISL01	0.0000
HISL08	0.0000	HISL08	0.0000

11.3.1 Model 3: CB and Predictor Importance of the Land Features and Socio-demographic Variables

The top 15 predictors are the following: BLD, COM02, FP, AREA, MORT02, HTEN02, ELEV, Language: English only (LANG02), YCOR, PPDN, NZKT, NZSW, SLPE, NZMN, and Income and Benefits: $150,000 to $199,999 (INC10). The median R^2 is 0.74. Running 50 simulations of the top 15 predictors only, the median R^2 is found to be 0.78. Box plots of the variables of each simulation are shown in Fig C.1 and listed in Table C.3.

Abbreviation	Variable	Percent Importance
BLD	Number of buildings	23.20
COM02	COMMUTING TO WORK - Workers 16 years and over - Car, truck, or van -- drove alone	10.77
FP	Sum of the building footprints	9.60
AREA	Area	7.22
MORT02	MORTGAGE STATUS - Owner-occupied units - Housing units with a mortgage	6.43
ELEV	Mean elevation	5.53
HTEN02	HOUSING TENURE - Occupied housing units - Owner-occupied	5.15
SLPE	Slope - mean percent rise	4.84
YCOR	Centroid of y coordinate	4.56
NZMN	Mean of hourly precipitation (non-zero values)	4.47
NZSW	Skewness of hourly precipitation (non-zero values)	4.36
PPDN	Population Density	4.00
NZKT	Kurtosis of hourly precipitation (non-zero values)	3.71
LANG02	LANGUAGE SPOKEN AT HOME - Population 5 years and over - English only	3.40
INC10	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) - Total households - $150,000 to $199,999	2.82
		R^2 is 0.78

Table C.3 The median R^2 and relative importance values, resulting from 50 simulations of the RF regression for only the top 15 ranked predictors, with CB serving as the response variable (Model 3)

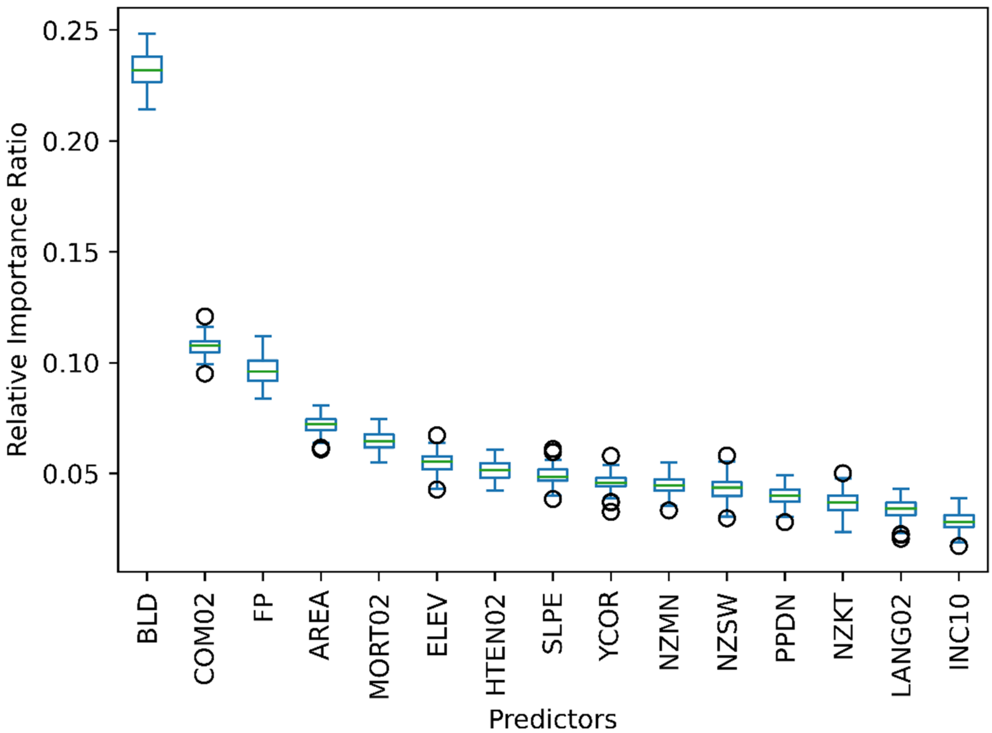

Fig C Box plots of the 50 RF simulations of the top 15 ranked variables only, with CB serving as the response. These 15 variables explain up to 78% of the spatial variability (The R^2 is 0.78). The expanded version of the acronyms is shown in Table C.1.

Of the top 15 predictors, five categories were socio-demographic. The COM02, MORT02, and HTEN02 were categories also found among the top predictors when SF served as the response. Then, there were also two new categories: Language and Income and Benefits. The Language Spoken at Home variable differentiates between those who speak only English in the home and those who do not. The Income and Benefits category discerns between the following annual incomes: less than $10,000, $10,000 to $14,999, $15,000 to $24,999, $25,000 to $34,999, $35,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $149,999, $150,000 to $199,999, $200,000 or more.

By the visualization of the top predictor categories and corresponding variables, it is seen that Model 1 (street flooding reports as the response variable) and Model 3 (catch basin reports as the response variable) share similar influences. Specifically, 10 of the top 15 predictors appear in both models. This may be expected, since street flooding reports and catch basin clogging reports occur during rain events; in addition, catch basin clogs have been shown to be a causal factor for street flooding. Thus, a location experiencing catch basin clogging may also be experiencing street flooding.

11.4 Appendix D

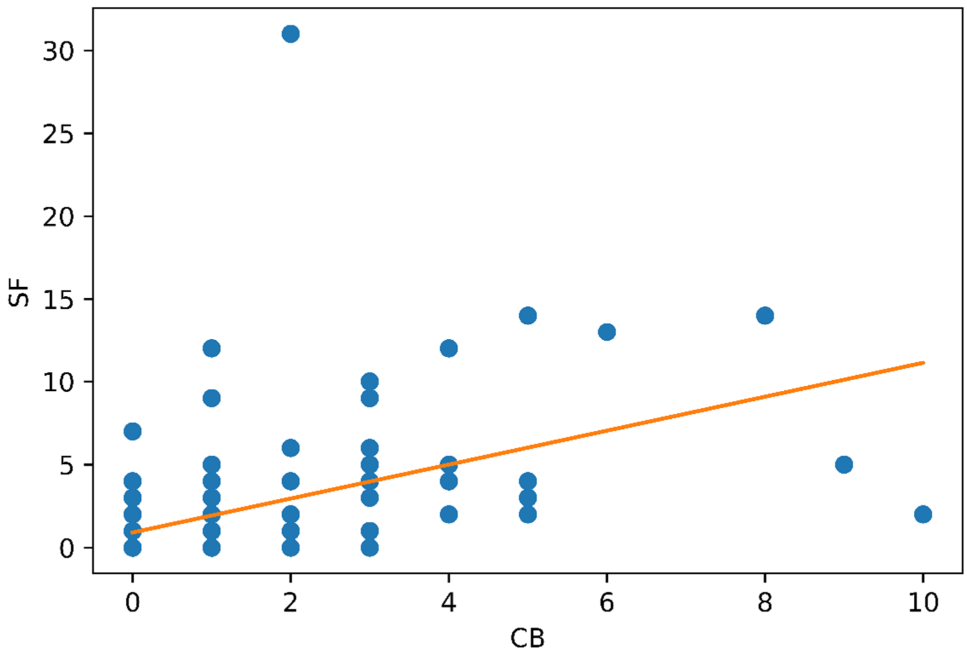

Fig D Scatter plots depicting the non-linear relationship of the CB and SF complaints on September 1, 2021, the day of the NYC urban flooding event by post tropical depression Ida. Each point represents a zip code, and the orange line represents the fitted regression line.