Understanding New York City Street Flooding through 311 Complaints

Candace Agonafir[ab*], Alejandra Ramirez Pabon, Tarendra Lakhankar, Reza Khanbilvardi[ab], 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) affiliation for all authors

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

Abstract

Street flooding is problematic in urban areas, where impervious surfaces, such as concrete, brick, and asphalt prevail, impeding the infiltration of water into the ground. During rain events, water ponds and rise to levels that cause considerable economic damage and physical harm. Previous urban flood studies and models have evaluated the factors contributing to street flooding, such as precipitation, slope, elevation, and the drainage network. Yet, due to the complexity of the interconnectedness of these factors and lack of available data, difficulty remains in ascertaining the localized areas prone to and experiencing street flooding. Thus, residents and city management of problem areas are unaware and unable to prepare for street flooding events. This study presents an evaluation of New York City's 311 street flooding reports, via an inference model, as a way to detect the zip codes where street flooding is prevalent. The potential explanatory variables for street flooding complaints were precipitation amounts and 311 sewer back up (water arising from home drains as a result of rainfall), manhole overflow (water arising from manhole covers on the street) and catch basin (a clogged basin preventing rainwater from entering storm drains) complaints. Using Stage IV radar precipitation data and 311 sewer reports, spanning a 10-year period, a Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis, with an embedded Zero-Inflation model is used to detect the variables statistically significant as predictors of flood complaint counts, specific to each zip code. The model is also tested using an Out-of-Sample prediction scheme by training it with the detected explanatory variables. Precipitation was found to be a predictor in 81% of the zip codes. For the infrastructural variables, manhole overflow complaints were significant to street flood complaints in 21% of the zip codes, back up complaints were significant in 41% of the zip codes, and catch basin complaints were significant in 47% of the zip codes. Thus, for an appreciable number of zip codes, infrastructural complaints were found to be predictors of street flooding complaints. This is the first study of its kind to investigate the infrastructural contributions of street flooding by 311 analysis, thereby identifying factors of street flooding, aside from precipitation. Leading contributions of the study include the demonstration of infrastructural impact towards the occurrence of street flooding and also the circumscription to the zip code and borough levels, allowing for tailored preventative actions in critical areas.

Highlights

- Crowd-sourced data (311 street flooding complaints) were analyzed to detect key explanatory variables that explain New York City's street flooding complaints.
- Catch Basin and Sewer Back-Up variables were shown as predictors in over 40% of the tested zip codes, revealing the adverse contributions of the drainage network towards street flooding occurrence in New York City.
- Some boroughs have a low frequency of reports; yet, in those areas, street flooding complaints are strongly influenced by small increases in the predictors.

1 Introduction

Flooding events result in multiple fatalities and considerable property losses each year. Particularly, within the urban environment, the effects are pronounced. Urban watersheds, lined with impervious surfaces, such as concrete, asphalt, and stone, have a limited amount of infiltration and recharge during heavy rainfall; thus, surface flow dominates the hydrological response (Serrano, 2010). Also, as the drainage system becomes overwhelmed, water overflows as runoff, and pluvial flooding, or what is commonly known as street flooding, occurs. Furthermore, as urban areas are densely populated, the consequences of flooding are oftentimes more severe than those of coastal or tidal flooding events. Indeed, for a given storm, more economic damage and injuries have been shown to occur in urban areas, as opposed to rural areas (Sharif, Yates, Roberts, & Mueller, 2006). For example, the National Weather Service (NWS) reported that, in 2014, a single, urban flooding event in Detroit, Michigan, resulted in $1.8 billion of direct damages, representing 60% of the total flood damages for that year in the United States (NWS, 2020a). In addition, in a study by the Chicago's Center for Neighborhood Technology (CNT), the economic costs of urban flooding for the densely populated area of Cook County, Illinois, totaled more than $773 million over a five-year period. (CNT, 2020). Thus, due to the unique physical and social characteristics of an urban area, flooding has acute impact.

The modeling of street flooding has the potential to reduce the economic and social effects of severe storms in urban developments. Specifically, the estimation and projection of flooded

69 areas has great benefit, as it allows for the implementation of early warnings, which, in turn,
70 provides people with the opportunity to take shelter and perform preventative measures. In recent
71 years, urban models, based on a variety of methodologies, including cellular automata, image
72 processing, and physically based systems have been introduced (Guidolin et al., 2016, Lo, Wu,
73 Lin, & Hsu, 2015). Generally, these models include analyses of rainfall, infiltration, and the sewer
74 system. In urban flood simulations, it is common to evaluate extended surcharge and other aspects
75 of the drainage network by dual drainage modeling, which incorporates the interaction between
76 surface flow and the sewer flow of surcharged sewer systems (Djordjević, Prodanović, &
77 Maksimović, 1999). Distinctly, extended surcharge occurs when water is held under pressure
78 within a sewer system during a rain event, thereby preventing the surface water to enter the
79 drainage system or causing the water from the drainage system to escape to the surface (Schmitt,
80 Thomas, & Ettrich, 2004). Within the United States, the most widely used flood forecasting model
81 is the Flash Flood Guidance of the NWS, which offers a deterministic, physically-based,
82 hydrologic model, utilizing real-time radar and satellite precipitation estimates (Ntelekos,
83 Georgakakos, & Krajewski, 2006, World Meteorological Organization, 2020). Thus, as shown,
84 there are various models, and the ongoing research demonstrates the interest of emergency
85 management to produce an effective model, customized to the metropolitan area.

86 While the production of urban flood models, particularly physically-based models, is in
87 continuum, nonetheless, there are obstacles. For instance, the NWS model may forecast floods;
88 yet it does not consider urban factors. Also, the NWS and other models incorporate rainfall;
89 however, they do not include some infrastructural factors, such as back up flooding. Moreover,
90 with the building of a flood forecasting model, other hurdles, including cost effectiveness and data
91 availability present. Specifically, in older metropolitan cities, the design of the drainage system is

92 oftentimes unavailable (Al-Suhili, Cullen, & Khanbilvardi, 2019). For instance, Zahura et al found
93 that physics-based models, such as TUFLOW, also suffered impairments by insufficient drainage
94 data (Zahura et al., 2020). In addition, urban flood forecasting models (including flash flood
95 models) have the distinct challenge with the validation of accuracy. For example, flash floods are
96 often caused by severe storms occurring only within six hours of rainfall (NWS, 2020b); hence,
97 there is a difficulty in quantifying measurements in the brief timespan. Urban flood forecasting
98 models, at timescales longer than that of the flash floods, also have limitations as they might not
99 be benchmarked with real observations. Consequently, there is a hinderance in the comparison of
100 model results with the physical system. Therefore, there is a need for a low-cost, empirical/data-
101 driven analysis which would illuminate the exact urban areas flooded during a rain event, in
102 addition to providing insight into the specific sewer infrastructure issues within those areas.

103 Accounts by persons directly experiencing street flooding may resolve some of the issues
104 and provide clarity into the occurrence, extent and driving mechanisms of street flooding particular
105 to an urban place. In New York City (NYC), there is a platform, referred to as 311, where residents,
106 business owners, and visitors are able to file issue reports to the NYC government, via phone,
107 website, or social media (Minkoff, 2015). For instance, an observer who notices street flooding
108 may enter the NYC 311 website and input the description, nature, address, and date and time of
109 the occurrence. These filings by New Yorkers are invaluable, as the 311 complaints, via catch
110 basin, manhole, and sewer back up reports, offer infrastructural insight, into the response of NYC
111 sewer system, of which available drainage data is insufficient. Moreover, street flooding reports
112 may serve an additional benefit. As time, date, and exact location of a complaint is listed, the 311
113 street flooding complaints may serve as tool for urban flooding model validation, as a model's

114 prediction of flooding in an area may be supported by an analysis of the local reports. Thus, the
115 data provided by 311 is a way to understand the causes and effects of street flooding.

116 This study presents an inference model, which highlights the key climate and
117 infrastructural variables that govern street floods in NYC. Of NYC, the 311 complaints are
118 aggregated over seven days (weekly time-scale) and to the zip code level. Street Flooding reports
119 are taken as the response variable, whereas Precipitation amounts, Sewer Back-Up, Manhole
120 Overflow, and Catch Basin reports serve as predictors or explanatory variables. Utilizing the Least
121 Absolute Shrinkage and Selection Operator (LASSO) regression analysis (Tibshirani, 1996), with
122 an embedded Zero-Inflation (ZI) model, per zip code, the variables effecting street flooding
123 complaints are selected. By identifying the climate and infrastructural issues, areas prone to street
124 flooding and their particular vulnerabilities are revealed, thereby providing direction and clarity
125 for city management and forecasters. Furthermore, such an analysis complements the physical
126 modeling endeavors and provides tools of validation.

127 There have been a few studies, of which crowd sourcing was applied in flood analyses. In
128 one such paper, Sadler et al., flood severity had been analyzed and the data reported by residents
129 and individual observers was utilized to provide an inference model. As Sadler et al. delved
130 extensively into environmental factors, such as water table level and rainfall intensity (Sadler et
131 al., 2018), this study differs by reviewing infrastructural factors, such as issues involving the
132 drainage network and external catch basins. Additionally, there have also been flood analyses,
133 which have specifically used the NYC 311 format. For instance, Kelleher and McPhillips
134 employed NYC 311 complaints to explore the relationships between topographic indices and
135 pluvial flooding (Kelleher & McPhillips, 2020). While the study highlights the value of citizen
136 reports as a validation tool, it, however, does not analyze 311 street flood complaints in regards to

137 climatic or drainage sources. In another study by Smith and Rodriguez, street flooding complaints
138 were used to investigate topographic issues, in addition to serving as a validation method for a
139 proposed rainfall dataset (Smith & Rodriguez, 2017). Yet, as only street flooding and highway
140 flooding complaints were compiled, the infrastructural related 311 complaints were not assessed.
141 In contrast to previous research, this study is unique in its evaluation of sewer-related issues and
142 their effect on street flooding.

143 The paper is outlined in the following manner. In Section 2, the study area and data
144 processing are described. Relative information on NYC is set forth, with a focus on the climatic
145 and topographic elements, population density, borough and Sewershed delineations, and drainage
146 networks. Next, the data collection of the 311 complaints and radar precipitation is discussed,
147 along with the tools and methods involved with the pre-processing. Section 3 offers the
148 methodology of the analysis. There is an evaluation of the quantity and frequency of complaints
149 at the zip code and borough levels. In the methodology section, the Lasso ZI is introduced as well,
150 along with the Negative Binomial Generalized Linear Regression Model (nbGLM) ZI, where the
151 prior identifies the infrastructural and climatic predictors, which feeds into the latter for Out-of-
152 Sample (OOS) predictions. In Section 4, the results of the model are presented, including the
153 mapping and tabulations of coefficients, variability, and error determinations and their
154 implications are discussed and interpreted. Finally, in Section 6, summary and major conclusions
155 are presented.

2 Study Area and Data

2.1 Study Area

NYC is located in the northeastern United States, at the coast of the Atlantic Ocean. It is markedly impervious and populous, which makes it an ideal study area for urban flooding. Spanning only 800 square kilometers, NYC has the highest population of any U.S. city, and it also has the greatest density (United States Census Bureau, 2012). Moreover, dissimilar to other U.S. cities, NYC is comprised of five boroughs (each representing a separate county): Queens, Brooklyn, Manhattan, Bronx, and Staten Island. Of the boroughs, Queens and Brooklyn have the highest populations, at approximately 2,200,000 and 2,500,000 people, respectively; Manhattan, with approximately 1,500,000 residents, has the highest population density; Bronx has approximately 1,300,000 residents; and, Staten Island is the least populous at 470,000 residents (United States Census Bureau, 2020). In regards to ground topography, approximately 72% of the land area of NYC is covered with impervious surfaces (City of New York, 2020a). A map of percentage impervious surfaces is shown in Figure 1a.

Concerning the climate of NYC, the classification is humid subtropical (NWS, 2020c), according to Köppen-Geiger Climate Subdivisions. The mean daily temperature is 13 °C, and the yearly rainfall in NYC is roughly 1270 millimeters (NWS, 2020d). Annually, the mean number of days with precipitation of 0.254 millimeters or higher is 120 days (National Oceanic and Atmospheric Administration, 2020a), and the mean number of days with precipitation of 25.4 millimeters or higher is 13-14 days (State of New York, 2020). In New York and areas of the Northeast, annual precipitation is uniformly distributed (Petersen, Devineni, & Sankarasubramanian, 2012). According to the New York State Climate Hazards Profile, NYC has experienced between 90-102 severe storms between the years 1960 through 2014, and the

179 subsequent costs ranged between $4 to $17 million (State of New York, 2020). In addition, due to
180 climate change, it is projected that precipitation extremes are expected to increase in the future
181 (González et al., 2019).

182 With respect to infrastructure, the catch basins of NYC connect the storm water to the
183 underground sewer system. A map of the number of catch basins per square kilometer is shown in
184 Figure 1b. Of the sewer connections, there are two types of drainage systems in NYC: Combined
185 Sewer System and Separate Storm Sewer System. The Separate Storm Sewer System uses separate
186 pipes: one pipe to carry wastewater to the wastewater plant, and a different pipe to carry
187 stormwater to the waterways (City of New York, 2020b). Most of NYC is comprised of the
188 Combined Sewer System, which uses a single pipe to transport both wastewater and stormwater
189 to a wastewater treatment plant (City of New York, 2020b). Servicing drainage areas, ranging from
190 13 to 102 square kilometers, there are fourteen wastewater treatment plants, which are also known
191 as *Sewersheds* (City of New York, 2020c). In addition, for the Combined Sewer System, when
192 there is heavy rainfall and capacity is exceeded, overflows occur, and a portion of the water
193 discharges to a Combined Sewer Outfall and enters a waterway (State of New York, 2020).

194 **2.2 NYC 311 Platform**

195 The NYC 311 sewer complaints data may be accessed via the NYC Open Data website:
196 data.cityofnewyork.us, where data is available from January 1, 2010 onwards. The complaints are
197 geocoded with the latitude and longitude of the location from where the complainant had stated
198 the issue had taken place. The date and time the complaints are also recorded. Through 311, a
199 person may file a complaint and categorize sewer complaint as follows: Street Flooding (SF), to
200 report flooding or ponding on a street; Sewer Back-Up (BU), to report, during heavy rainfall or
201 flooding, water arising from a toilet, sink drain or bathtub drain; Manhole Overflow (MO), to

report a manhole overflowing with water or sewage; or Catch Basin (CB), to report a clogged or damaged Catch Basin. For sewer back-ups, it shows a relationship between the private drains and the public sewer system, as back up flooding occurs when either the height of the water in the public pipes are greater than that of the gravity inlets inside the private property or when the inlet level of the storm drains are below the water level of the sewer (Schmitt, Thomas, & Ettrich, 2004). Regarding manhole issues, the overflowing of a manhole signifies surcharge, as water from the sewer system has travelled to the surface; thus, MO complaints may be indicative of infrastructural issues. Lastly, as catch basins are the grates allowing for the collection of storm water, CB reports provide useful knowledge to street flooding behavior. If catch basins are blocked or malformed in certain areas, surface water level increases, and this may be indicative of city maintenance problems.

2.3 Radar Data

The National Center for Atmospheric Research (NCAR)/Earth Observing Laboratory (EOL) website offers NCEP/EMC 4KM Gridded Data (GRIB) Stage IV datasets, where hourly, 6-hour, 12-hour, 24-hour totals of millimeter precipitation amounts are available from years 2001 through 2020. As the Stage IV data is unable to adjust for severe snow events, the data in the northeastern United states include only rainfall data (Hamidi et al., 2017). From the EOL website, 24-hour radar precipitation data, from years 2010 through 2019 were ordered. The Thiessen Polygon Method (Viessman & Lewis, 2003) was employed, with each radar point as center, to aggregate the gridded radar precipitation data available at the 4 km by 4 km resolution to the zip code resolution. With the use of the Thiessen Polygon method of Arc GIS Pro, a weighted average of radar points within a zip code boundary was calculated. Then, the rainfall amount per zip code

224 was determined using this weighted average. The data was finally aggregated to the weekly time
225 scale, i.e., total precipitation (in millimeters) per week (PRCP).

226 **2.4 Data Collection, Processing, and Preliminary Analysis**

227 Sewer Complaints data using 311 reports, from January 1, 2010 through December 2019,
228 were downloaded from the NYC Open Data, government website. The data was geo-aggregated
229 to the zip code level and only the issues relating to street flooding were extracted. In addition, to
230 account for possible lags in the occurrence of an event and the report of the issue, weekly sums of
231 each complaint were calculated. A reason for lags is that a person may take time to report an issue.
232 This may be especially true in urban areas, where warm season rainstorms producing short-
233 duration, heavy rainfall, oftentimes, take place in the evenings (González et al., 2019). Also, there
234 may be lags between the rain event and the occurrence of street flooding, such as, for instance,
235 when the drainage system becomes more overwhelmed with debris as time passes. Since the exact
236 detection of the lag that measures the difference between the time of the event(s) and the time of
237 the complaint(s) may be arduous, for simplicity, a weekly timescale (Sunday to Saturday) was
238 decided as the unit of temporal aggregation for all the variables. It is assumed that a week is not
239 far removed to have lost the influence of precipitation resulting in street flooding complaints. The
240 same is true for infrastructure complaints where the infrastructure complaints within a week are
241 assumed the possible antecedents of the street flooding complaints that week.

242 Another measure taken was to ensure that the same complainant was not reporting a
243 specific location repeatedly. By the mechanism of the 311 website, a complainant may report the
244 same location more than once in a day. To see whether a location was reported more than once in
245 a day, the SF, MO, BU, and CB complaints over the ten-year period were processed for their
246 uniqueness. The 311 data lists each complaint as a row, containing latitude and longitude location

247 coordinates. Only the unique location coordinates were retained in this study. Of the raw 311 data,
248 from January 1, 2010 through December 31, 2019, there were 25,574 SF, 6,042 MO, 137,974 BU,
249 and 85,607 CB total collective reports, and it was determined that 25,378 (99.2%), 5,687 (94.1%),
250 128,751 (93.3%), and 82,191 (96.0%) were unique, respectively.

251 Zip code, borough, and catch basin shapefiles were downloaded from NYC Open Data and
252 processed via ArcGIS Pro. After all data was processed, 174 zip codes, 530 weeks of precipitation
253 totals and 311 SF, BU, CB, MO complaint totals, over the ten-year period, from January 1, 2010
254 through December 31, 2019, were used for analysis.

255 Before the development of the model, a complaint frequency analysis was conducted. Per
256 zip code, the number of SF complaints over 10 years were computed and examined (Figure 2).
257 The median of total complaints per zip code was 87, with 1300 being the max and zero being the
258 minimum. The histogram (Figure 2a) shows that the majority of zip codes reported under 200
259 complaints during the 10-year period (136 zip codes, 78%). To illustrate the zip codes most
260 frequently reporting SF complaints, the average of the total complaint for all zip codes were taken
261 (average total complaints = 139), and the zip codes with a total complaint value greater than the
262 average of 139 complaints were identified. Figure 2b presents a map of the total complaints per
263 zip code where the zip codes that have total complaints greater than the average total complaints
264 are highlighted. The illustration shows Staten Island, lower Brooklyn, and Queens as having the
265 highest frequencies of SF reports. Per borough, the number of complaints per 10,000 people are
266 98.4, 44.5, 24.6, 15.8, and 13.3 for Staten Island, Queens, Brooklyn, Manhattan, and Bronx,
267 respectively.

3 Methodology

The NYC Department of Environmental Protection identifies Increased Precipitation, Blocked Catch Basin Grates, and Surcharged Sewers [leading to Sewer Back Ups] as major causes of flooding in NYC (City of New York, 2020d). With a yearly average precipitation of 1270 millimeters, NYC experiences significant precipitation through the year, with little intra-annual variations. However, there is a considerable spatial variation within NYC (Hamidi et al., 2017), which may result in localized street flooding. Blocked catch basin grates may also lead to street flooding. Intense storms may push leaves and litter onto catch basins, where they could mold into mats and obstruct the basins. Blocked catch basins prevent rainwater from entering the storm sewer, thereby causing street flooding. Frequently, during intense rainfall events, the combined volume of stormwater and wastewater exceeds the sewer system's capacity. Under such circumstances, the excess stormwater remains in the streets leading to flooding.

The hypothesis of this study is that the climatic and infrastructural issues are statistically significant predictors of the response, 311 SF complaints. Precipitation, the climatic feature, is the primary cause of flooding. In addition, sewer surcharge, as indicated by back up and manhole overflow issues, or the blockage of stormwater drains by catch basins, also contribute to street flooding. For variable identification, a LASSO ZI, which imposes a penalty function, cancelling out the coefficients of less important variables, was implemented. The LASSO method shares the usual model assumptions concerning the nature of the relationship between response variable and the explanatory variables, but adds an important L_1 constraint to the regression coefficients in least squares optimization. The result is the inevitable shrinkage of certain coefficients to zero, allowing the LASSO technique to enjoy advantageous properties of ridge regression and best subset selection (Tibshirani, 1996; Hastie, Tibshirani and Friedman, 2001).

Then, a ZI, generalized linear modeling framework was used to perform OOS predictions, using an eight-two-year training and testing data set, as to show the variability in the SF complaints using PRCP, CB complaints, BU complaints, and MO complaints. Since the SF complaints data is discrete, and since the counts per week are being measured, a Negative Binomial model was employed as the link function. The Negative Binomial model is a generalization of the Poisson regression models that accounts for overdispersion (Lawless, 1987).

For variable selection, the Multicollinearity-adjusted Adaptive LASSO for Zero-inflated Count Regression (AMAZonn) package in R was used. The algorithm allows for the implementation of LASSO, with a ZI nbGLM element (Mallick, 2018). By shrinking the coefficients of the predictors or tuning the coefficients to zero, LASSO creates a subset of the predictors that have the most effect on the response, allowing for more interpretable results and higher prediction accuracy (Tibshirani, 1996), and ZI models accommodate excess zeroes, of which the nbGLM cannot, by providing a two-component model, a point mass at zero and a Poisson, geometric, or negative binomial (Zeileis, Kleiber, & Jackman, 2008). As the 311 count data is discrete, and there are many weeks with zero complaints, the LASSO with a ZI nbGLM was appropriate.

The nbGLM part of the model, with *y* as the response variable with the four predictors for each zip code *i*, is shown here:

$$y_{it} \sim NB(p_{it}, r_i) \ldots (1)$$

where,

$$p_{it} = \frac{r_i}{r_i + \lambda_{it}} \ldots (2)$$

$$\lambda_{it} = e^{[\beta_i^0 + \beta_i^1 * PRECIP_{it} + \beta_i^2 * CB_{it} + \beta_i^3 * BU_{it} + \beta_i^4 * MO_{it}]} \ldots (3)$$

Equation (1) shows that the weekly aggregated street flooding complaints in each zip code (y_{it}) is modeled as a Negative Binomial distribution with a success parameter (p_{it}) and an overdispersion parameter (r_i). The success parameter (p_{it}) relates to the rate of occurrence (λ_{it}) [Equation (2)], which is informed by a regression on the precipitation ($PRECIP_{it}$) and infrastructure covariates ($CB_{it}, BU_{it}, MO_{it}$) [Equation (3)]. β_i^0 is the regression intercept for zip code i, and $\beta_i^1, \beta_i^2, \beta_i^3, \beta_i^4$ are the regression slopes representing the sensitivity of the street flooding complaints to precipitation ($PRECIP_{it}$), catch basin complaints (CB_{it}), sewer back up complaints (BU_{it}), and manhole overflow complaints (MO_{it}), respectively. These model parameters are estimated using a maximum likelihood approach in R version 4.0.4 (Friedman et al., 2010).

The explained variance (pseudo-R^2) of the nbGLM, which is estimated as $1 - \left(\frac{L(0)}{L(\boldsymbol{\beta})}\right)^{2/n}$, where $\frac{L(0)}{L(\boldsymbol{\beta})}$ is the ratio of the likelihood of the null model to the fitted model and n is the sample size, demonstrates the extent to which the model explains the variability in the response (Cox and Snell, 1989). As the 311 complaint data was discrete, the fit index for a redefined pseudo-R^2, proposed by Nagelkerke (Nagelkerke, 1991), was utilized. This redefined measure normalizes the model pseudo-R^2 to the maximum possible achievable using the likelihood ratio estimate.

For the OOS predictions, eight years were used as training data, and two years as testing data. Using a k-fold cross validation technique, the training data consisted of eight years of the SF, BU, CB, MO, and PRCP weekly data, with the remaining two years serving as the testing set. The years were randomly shuffled, such that the training set may consist of a different eight grouping of years between 2010 through 2019 and a subsequent different two year grouping of the testing set. Using the Lasso selected variables, the model is "trained" by the influence of the predictors towards the outcome, SF, during the eight [not necessarily consecutive] years. Predictions of SF,

based on the observed predictors for the two years, are then conducted using the trained nbGLM ZI model (For the nine zip codes where LASSO did not select a significant predictor, a standard nbGLM is utilized, without LASSO selection, to obtain predicted values). The predicted SF values are then compared to the actual SF Values. For each random selection of training and testing sets, simulations were run 100 times, and the mean arctangent absolute percentage error (MAAPE) values were determined per zip code. MAAPE accommodates data with zero values by the application of slope as an angle, as opposed to slope as a ratio (Kim and Kim, 2016):

$$MAAPE = \frac{1}{106}\sum_{t=1}^{106} arctan\left(\frac{O_t - P_t}{O_t}\right) \; for \; t = 1, 2, \ldots, 106 \ldots (4)$$

O represents the observed SF weekly complaints for the two-year period (106 weeks), and P represents the predicted SF values. By the equation, it is seen that a closer value between the observed and predicted would result in a value closer to zero, and a larger difference between the observed and predicted would result in a value converging to $\frac{\pi}{2}$ radians.

In summary, the modeling framework has the following steps:

1. For each zip code, statistically significant predictors are identified by the use of the multicollinearity-adjusted adaptive LASSO, implemented with the ZI nbGLM.
2. The statistically significant predictors by zip code are reported as the most important features for understanding street floods in that zip code.
3. A ZI nbGLM is trained using the LASSO inferred variables for each zip code, and the model's efficacy is tested using OOS predictions against the held-out data.

This final step provides additional robustness to the model and its selection.

4 Results and Discussion

4.1 The Circumstance of NYC Street Flooding

By citizen imported data, this study first maps the locations where street flooding is often reported. When examining the total SF reports over the 10-year period, the presence of flooding is highest in Staten Island, lower Brooklyn, and various zip codes in Queens. The complaints are localized to the zip code level to allow for a tailored insight into the areas where street flooding occurs the most, as this would be necessary for flood forecasting at the neighborhood or street level. As each borough represents a separate county within NYC, this study included a localization to the borough level, as well. In addition, an examination of the reports at the broader borough level is also beneficial to stakeholders and policy makers, as borough boards are able to create bylaws and plans. In this consideration, Staten Island and Queens are of special interest. Per 10,000 residents, Staten Island has the most complaints, which is roughly double the complaints of Queens, the second highest frequency borough. Likewise, Queens has almost twice the complaints of Brooklyn, which follows in third. Moreover, as a 311 complaint, by its nature, is citizen reported, street flooding is not only occurring, but is also adversely felt by the residents, especially those in Staten Island and Queens.

4.2 Response to Predictors and Their Significance

The regression analysis provides a selection of predictors and the degree of their influence. In Figure 3, the zip code level significant explanatory variables were based on the inference of the regression coefficients ($\beta_i^1, \beta_i^2, \beta_i^3, \beta_i^4$). The strength of the association, ($e^{\beta_i} - 1$) for infrastructure and precipitation covariates, are expressed as percentage change in the expected weekly counts per unit change in the explanatory variable, and it is shown in the graduated color scheme. The zip codes designated in white did not have the variable selected as predictor by LASSO. The intercept

from the model (β_i^0) for each zip code is also shown in Figure 3e (plotted as $e^{\beta_i^0}$). As expected, there is similarity to the frequency map, as the intercept exhibits an upward shift with more complaints. Thus, insight into the behavior of the predictors is gained by the regression coefficients.

The spatial variability of the predictors is also observed. There was a total of 165 zip codes of the 174 zip codes in the study, where at least one predictor was selected by LASSO. PRCP was selected in 141 zip codes, of which 55, 12, 28, 20, and 26 zip codes were located in Queens, Staten Island, Brooklyn, Bronx, and Manhattan, respectively. BU was selected in 72 zip codes, of which 29, 6, 17, 9, and 11 zip codes were located in Queens, Staten Island, Brooklyn, Bronx, and Manhattan, respectively. CB was selected in 82 zip codes, of which 25, 9, 20, 8, and 20 zip codes were located in Queens, Staten Island, Brooklyn, Bronx, and Manhattan, respectively. MO was selected in 37 zip codes, of which 17, 2, 4, 5, and 9 zip codes were located in Queens, Staten Island, Brooklyn, Bronx, and Manhattan, respectively. Of the variables, PRCP was an explanatory variable in the most zip codes, followed by CB. BU is the third most represented explanatory variable. Lastly, MO is shown as an explanatory variable in the least amount of zip codes. Thus, while climatic and infrastructural variability have high selection, there are also notable differences among zip codes.

To further examine the spatial variability of the boroughs, each selected predictor's breakdown by borough is determined. In Table 1, for each predictor, where significance is found, the percent of zip codes in each borough is shown. In addition, Table 1 shows the ratio of the mean exponent of the β of each selected predictor of borough to the mean exponent of the β for NYC as total (BT Ratio) - a measure to understand the expected sensitivity of a borough relative to the expected sensitivity of NYC for each of the explanatory variables. A BT ratio greater than 1

401 signifies that the borough experiences a stronger reaction (greater increase in SF complaints), when
402 the LASSO selected predictor (either CB, BU, MO, or PRCP) experiences an increase in
403 complaints [or, in the case of PRCP, amounts], than that of NYC on average. A ratio lower than 1
404 signifies that the borough experiences a weaker reaction. By the table, the selected predictor and
405 strength of association is shown at the borough level and compared to the overall findings of NYC.

406 Plausibly, SF complaints may not be a comprehensive portrayal of the occurrence of street
407 flooding in NYC, as certain zip codes or boroughs may have residents with greater proclivities
408 towards addressing concerns. Yet, the selection of the predictor, PRCP, in 82% of the zip codes
409 (Table 1) demonstrate that, in the majority of NYC zip codes, the SF reports are consistent with
410 and heavily affected by rain events. In addition, the LASSO selection of the other predictors as
411 affecting SF reports further strengthens the validity of the 311 platform as an accurate portrayal
412 rainfall occurrence and effects. If reports were being made haphazardly, a connection between an
413 infrastructural element and street flooding would not be found by LASSO. Therefore, while there
414 may be additional factors affecting residents' complaints, there is sufficient accuracy in the 311
415 complaint filings, as the connection between the predictors and SF reporting, found by the model,
416 further validify the platform.

417 **4.3 Analysis of Model Parameters**

418 An analysis of model parameters also provide insight into the different occurrences among
419 boroughs. When looking at the analysis, it shows that, although there are areas with a high
420 frequency of SF reports, these areas do not necessarily have the greatest rate of SF report increase
421 when its predictor experiences an increase. This lack of sharp increases in SF compared to the
422 increases in the LASSO selected variables (CB, BU, MO, or P), coupled with a high frequency of
423 complaints (indicating active engagement on the 311 platform), may signal a chronic problem in

those areas, of which the residents appear to experience street flooding during moderate conditions (due to low beta values), and subsequently, file more complaints. Indeed, this is evident, especially in Staten Island. Examining Figure 2b, 10 of 12 Staten Island zip codes have a high frequency of reports. Yet, when looking at the infrastructural variables of significance in Figures 3 b-d, none of the zip codes have beta percentages in the highest quantile (Table 1). Thus, while BU and CB, specifically, show significance in 50% and 75% of the Staten Island zip codes, respectively, an increase in those complaints do not trigger the greatest increase in SF, as compared to zip codes in other boroughs. Interestingly, one such borough is Manhattan. Manhattan has only two zip codes with total complaints slightly greater than the average total complaints for NYC in total, when looking at Figure 2b. Yet, for instance, in Figure 3c and Table 1, Manhattan has CB as predictor in 48% of the zip codes, where, at least, eight zip codes are ranked in the highest quantile group, based on sensitivity. It may be inferred that the residents are not reporting 311 complaints (specifically SF reports, as shown in the frequency analysis) excessively in Manhattan. However, when there is a CB report, SF reports are strongly influenced. This is apparent for BU in Manhattan, as well; and, in Bronx, CB and MO, with high BT ratios (Table 1), respectively, also behave in a similar manner to Manhattan. Finally, it can be seen that zip codes have different sensitivities, as shown in the Figure 3 maps. This also supports the notion that zip codes suffer from varying infrastructural issues at varying extents. When a predictor is selected, the parameter analysis provides information regarding the severity of the effect, and at this study's localized level (an average area of 2.75 square kilometers per zip code), problem areas are pinpointed.

4.4 Variable Importance

4.4.1 Catch Basin (CB)

Catch basin infrastructural issues are of noteworthiness, since they directly lead to street flooding if they are not working properly. Catch basins are also an external component of the

448 drainage network. Therefore, the public has direct access to the basins and are able to assist or
449 damage them. Consequently, an outreach effort by NYC to the residents may be of help. One such
450 partnership exists in Newark, NJ, where there is a program called Adopt a Catch Basin (City of
451 Newark, 2021). The program offers residents the opportunity to use an ArcGIS Solutions mapping
452 platform to select a catch basin to adopt; they care for the basin, cleaning and removing debris;
453 then, they are also encouraged to paint and decorate the basin (City of Newark, 2021).

454 In this study, CB was selected as a predictor in almost half of the NYC zip codes in total.
455 While, similar to the frequency trend, Staten Island had the highest percentage of zip codes, at
456 75%, where CB was selected as a predictor. Queens and Brooklyn followed, at 42% and 54%,
457 respectively. Finally, there were also many zip codes in Manhattan where CB was selected as a
458 predictor (48%), despite Manhattan having a low number of total complaints. Furthermore, in
459 Manhattan, the difference between zip codes with PRCP selected as a predictor (60%) to the
460 number with CB selected (48%) was smallest of the boroughs. It is possible to infer that the
461 contrast of model results from one borough, such as Manhattan to the others, highlights specific
462 issues within the zones. When looking at the map of impervious surface percentage (Figure 1a), it
463 is seen that Manhattan has the highest percentage of impervious surfaces. Thus, a possible theory
464 for CB in Manhattan having a high BT ratio and selection percentage is that the storm runoff may
465 be carrying trash into the stormwater drains, thereby clogging the catch basins. Specifically,
466 Manhattan has more active construction sites than any other borough (City of New York, 2020),
467 and waste from sites are a contributing factor to runoff debris in urban areas (Environmental
468 Protection Agency, 2003). Overall, for an infrastructural category, CB complaints were selected
469 as predictors in a large number of zip codes. This is an impactful finding, as it indicates that,

oftentimes, when one person observes and reports a street flooding event, there is another person observing and reporting water ponding from a clogged catch basin, within that time period.

4.4.2 Sewer Back-Ups (BU) and Manhole Overflows (MO)

Concerning BU, when looking at Figure 3b, there appears to be a noticeable shift inland, when comparing the areas to those of CB selected predictor, as shown in Figure 3c. For Bronx, the results were similar to the other boroughs in regards to BU selection; whereas, for CB, the Bronx had a much lower number of zip codes showing significance. Also, Manhattan appears to have the lowest BU issues. Aside from location, BU performed similarly to CB, with 41% of zip codes having the variable selected as a predictor. For zip codes experiencing explanatory power from a combined PRCP, with BU or MO issues, it signifies a chaotic condition, where it is not only raining and the streets are flooded, but internal drains are being overwhelmed and working in reverse order. An internal drainage issue may not be as easily remedied, as with catch basins, where maintenance and public awareness may have a positive effect; however, areas shown on the maps, where BU and MO issues are signified, should be investigated, monitored, or modeled, as it may facilitate long term planning improvements.

It has been theorized that a difference between the topographic wetness index concerning flood reports of Staten Island and Manhattan is due to the type of construction of the combined sewer overflow system in Manhattan, compared to that of the separate sewer system in Staten Island (Kelleher & McPhillips, 2020). However, with the inclusion of all boroughs, the results in this paper show that the zip codes in Brooklyn, which are mostly comprised of the combined sewer system, have back up issues as a predictor of street flooding in 46% of its zip codes. When reviewing the Open Sewer Atlas data (Open Sewer Atlas NYC, 2021), a web resource directed from the NYC Open Data website (City of New York, 2020e), 80% of the zip codes in this study

493 are within the combined sewer system. When reviewing the results of this study, 81% of the zip
494 codes with BU as a selector are located in a combined sewer system. Thus, there appears to be no
495 difference in NYC between the combined sewer system and separate sewer system in regards to
496 SF reporting. Finally, concerning manhole overflow complaints, there is significance in few zip
497 codes. As these areas are variously located throughout NYC, there is indication that the sewer
498 issues are area-specific. Localized to the manhole level, the mapping of these particular zip codes
499 would be of aid to city management in the investigation of issues within the internal drainage
500 network.

501 **4.4.3 Precipitation (PRCP)**

502 As expected, PRCP is the primary driver of street flooding. PRCP is shown as an
503 explanatory variable in 81% of the zip codes. Regarding the beta results, street flooding reports
504 respond greatest to changes in precipitation in Staten Island and Brooklyn. However, many zip
505 codes in Queens and Bronx also exhibit strong increases to street flooding complaints due to
506 increases in precipitation amounts. Figure 3a highlights the zip codes prone to dramatic increases
507 in street flooding; as thus, particularly in those areas, precautions may necessary to take in the
508 advent of a forecasted severe rain event. Our future modeling will include rainfall intensity, in
509 addition to duration, as they are key elements in flash flooding (NWS, 2020e).

510 **4.5 Explained Variability**

511 Pseudo-R^2 and MAAPE determinations were used for an understanding of variability. To
512 illustrate the dependence of explained variability on the number of complaints, pseudo-R^2 values
513 were mapped against total complaints in Figure 4. In addition, the pseudo-R^2 values are depicted
514 in Table 2 and mapped in Figure 5 to show an aspect of variability. As an additional measure of

515 variability, MAAPE values, determined from the observed and predicted values by the OOS
516 predictions, are illustrated in Figure 6.

517 **4.6 Model Limitations**

518 There were factors which appeared to affect the pseudo-R^2 and MAAPE values. The mean
519 pseudo-R^2, determined by the nbGLM was 0.14. Boroughs, such as Staten Island, Brooklyn, and
520 Queens, had pseudo-R^2 values greater than the mean, at 0.22, 0.16, and 0.15, respectively (Table
521 2). As Figure 4 highlights, pseudo-R^2 values trend greater when there are a higher number of SF
522 complaints. Similarly, it is seen that lower MAAPE values (lower errors), as shown in Figure 6,
523 occur in the zip codes with greater total of SF complaints. Thus, if the model were to be constructed
524 on a larger grid scale, or if there was more data on the street flooding complaints, the pseudo-R^2
525 values would increase, and the OOS predictions would have improved results. However, this study
526 sought a localized scale, as to identify problem areas. An additional insight gathered from the
527 increase in variability due to low complaints is that the promotion of crowd-sourced platforms is
528 important. This study was limited by date range. The 311 data is available from 2010 onwards,
529 and if the data had been collected earlier, there would have been more complaints. This research
530 may have also been limited by low resident participation. As the study indicates infrastructural
531 complaints, oftentimes, are in relation with SF complaints, increasing awareness to residents and
532 visitors of NYC, especially when there is forecasted precipitation, would facilitate modeling
533 endeavors. It is essential to not just model the capacity and include the locations of the drainage
534 network, but the assessment of the performance capability, or current conditions, of the network
535 needs presence. Thus, encouraging residents to file reports when a sewer related issue occurs will
536 be beneficial. As shown, the model experienced limitations due to the small scale; in addition, the
537 model would benefit by increased resident participation.

Future research could include a predictive model, of which the findings of this inference model will lend insight into. This future study could also include rainfall intensity, as it is a key element in runoff. Furthermore, as the data have been aggregated to weekly values, a study utilizing a smaller range may have certain benefits. Similarly, a sewershed aggregated or city aggregated analysis would enable the better incorporation of spatial covariates and provide insights about its spatial variability. An example of such city aggregated model is presented in the Appendix (Table A). The results of the nbGLM conducted for NYC in whole, including topographical elements, such as slope and elevation, in addition to population. CB, BU, and elevation were found to be statistically significant, and the pseudo-R^2 was found to be 0.53.

5 Summary and Conclusions

With the advent of social media and smart phones, crowd-sourcing has become an effective tool for scientists to access data, which would otherwise be difficult or impossible to obtain. This study has found insights regarding street flooding in NYC, one of the largest, metropolitan cities in the world. Moreover, as the analysis was performed at the zip code level, problem areas were identified, allowing for tailored interventions. While other papers have examined 311 street flooding reports, this is the first of its kind to include the infrastructural components of sewer back-ups, catch basin complications, and manhole overflows. These factors were investigated as explanatory variables of the response, street flooding reports. The data, which included radar precipitation estimates, were modeled via LASSO regression, with the potentially significant predictors fed into a negative binomial generalized regression model, where the resulting coefficients were analyzed, allowing for the interpretation of each predictor's significance. Finally, this paper conducted a geographical breakdown of total street flooding complaints, highlighting areas with the highest frequencies.

Major conclusions are drawn from this study. First, citizen reports are a valuable aid in detecting hydrological issues and offer first-hand insight into problematic areas. Second, the model illustrates that, while precipitation amounts are the largest factor in street flooding, back up and catch basin issues are also major contributors. This is not a comprehensive, predictive model; however, the pinpointed potential problem areas may give a starting point for agencies when installing sensors or Close Circuit Television footage. Finally, as infrastructural categories show significance, there is a potential for street flooding to be controlled in NYC by governmental actions. While there may not be actions to prevent the rainfall amounts, improving the internal and external components of the drainage network will reduce some of the physical and economic impacts of street flooding in metropolitan areas.

6 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.

7 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.

We are preparing a NOAA CESSERT server to host the data and the codes used in this study to which access will be given upon request.

8 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. We thank Joshua Rapp of NYC Emergency Management and Dave Radell, Nancy Furbush, and Nelson Vaz of the National Weather Service for providing valuable feedback on this work.

9 References

Al-Suhili, R., Cullen, C., & Khanbilvardi, R. (2019). An urban flash flood alert tool for megacities-Application for Manhattan, New York City, USA. *Hydrology*. https://doi.org/10.3390/HYDROLOGY6020056

CNT. The Prevalence and Cost of Urban Flooding. Retrieved from https://www.cnt.org/publications/the-prevalence-and-cost-of-urban-flooding. Accessed 30 Dec. 2020.

City of Newark. Adopt a Catch Basin. https://www.newarknj.gov/card/adopt-a-catch-basin. Accessed 25 May. 2021.

City of New York (2020). Essential Active Construction Sites. Retrieved from https://www1.nyc.gov/assets/buildings/html/essential-active-construction.html. Accessed 25 May. 2021.

City of New York a. Impact of NYW Bonds. Retrieved from https://www1.nyc.gov/site/nyw/investing-in-nyw-bonds/the-impact-of-investing.page. Accessed 30 Dec. 2020.

City of New York b. Combined Sewer Overflows. Retrieved from https://www1.nyc.gov/site/dep/water/combined-sewer-overflows.page. Accessed 30 Dec. 2020.

City of New York c. Wastewater Treatment System. Retrieved from https://www1.nyc.gov/site/dep/water/wastewater-treatment-plants.page. Accessed 30 Dec. 2020.

City of New York d. Flood Prevention. Retrieved from https://www1.nyc.gov/site/dep/environment/flood-prevention.page. Accessed 30 Dec. 2020.

City of New York e. NYC OpenData. Retrieved from https://opendata.cityofnewyork.us/projects/open-sewer-atlas-nyc/. Accessed 30 Dec. 2020.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data*. Chapman & Hall/CRC.

614 Djordjević, S., Prodanović, D., & Maksimović, Č. (1999). An approach to simulation of dual
615 drainage. In *Water Science and Technology*. https://doi.org/10.1016/S0273-1223(99)00221-
616 8

617 Environmental Protection Agency (2003). Protection Water Quality from Urban Runoff.
618 https://www3.epa.gov/npdes/pubs/nps_urban-facts_final.pdf. Accessed 30 May. 2021.

619 EOL. NCEP/EMC U.S. Gridded Multi-Sensor Precipitation (4 km). Retrieved from
620 https://data.eol.ucar.edu/dataset/21.089. Accessed 30 Dec. 2020.

621 González, J. E., Ortiz, L., Smith, B. K., Devineni, N., Colle, B., Booth, J. F., … Rosenzweig, C.
622 (2019). New York City Panel on Climate Change 2019 Report Chapter 2: New Methods for
623 Assessing Extreme Temperatures, Heavy Downpours, and Drought. *Annals of the New York*
624 *Academy of Sciences*, *1439*(1), 30–70. https://doi.org/10.1111/nyas.14007

625 Guidolin, M., Chen, A. S., Ghimire, B., Keedwell, E. C., Djordjević, S., & Savić, D. A. (2016). A
626 weighted cellular automata 2D inundation model for rapid flood analysis. *Environmental*
627 *Modelling and Software*. https://doi.org/10.1016/j.envsoft.2016.07.008

628 Friedman J, Hastie T, Tibshirani R (2010). Regularization Paths for Generalized Linear Models
629 via Coordinate Descent. Journal of Statistical Software, 33(1), 1–
630 22. https://www.jstatsoft.org/v33/i01/.

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

635 Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning: Data
636 Mining, Inference, and Prediction, Second Edition, New York, NY: Springer.

637 Kelleher, C., & McPhillips, L. (2020). Exploring the application of topographic indices in urban
638 areas as indicators of pluvial flooding locations. *Hydrological Processes*.
639 https://doi.org/10.1002/hyp.13628

640 Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent

641 demand forecasts. International Journal of Forecasting.

642 http://dx.doi.org/10.1016/j.ijforecast.2015.12.003

643 Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of*
644 *Statistics*. https://doi.org/10.2307/3314912

645 Lin, Y., & Mitchell, K. E. (2005). The NCEP stage II/IV hourly precipitation analyses:
646 Development and applications. *85th AMS Annual Meeting, American Meteorological Society*
647 *- Combined Preprints*, 1649–1652.

648 Lo, S. W., Wu, J. H., Lin, F. P., & Hsu, C. H. (2015). Visual sensing for urban flood monitoring.
649 *Sensors (Switzerland)*. https://doi.org/10.3390/s150820006

650 Mallick, H. (2018). *AMAZonn (A Multicollinearity-adjusted Adaptive LASSO for Zero-inflated*
651 *Count Regression)*. Retrieved from
652 https://github.com/himelmallick/AMAZonn/blob/master/LICENSE. Accessed 1 May 2021.

Minkoff, S. L. (2015). NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City. *Urban Affiars Review*, *52*(2), 211–246.

Mitra, A. (2016). *Fundamentals of Quality Control and Improvement* (Fourth). John Wiley & Sons.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*. https://doi.org/10.1093/biomet/78.3.691

Navarrete, Claudio Bustos & Soares, Filipa Coutinho (2020). dominanceanalysis: Dominance Analysis. R package version 2.0.0. https://CRAN.R-project.org/package=dominanceanalysis

National Oceanic and Atmospheric Administration a. Comparative Climatic Data. Retrieved from https://www.ncdc.noaa.gov/ghcn/comparative-climatic-data. Accessed 30 Dec. 2020.

National Oceanic and Atmospheric Administration b.National Stage IV QPE Product. Retrieved from https://www.emc.ncep.noaa.gov/mmb/SREF/pcpanl/stage4/. Accessed 30 Dec. 2020.

Ntelekos, A. A., Georgakakos, K. P., & Krajewski, W. F. (2006). On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM529.1

NWS a. United States Flood Loss Report – Water Year 2014. Retrieved from https://www.weather.gov/media/water/WY14 Flood Loss Summary.pdf. Accessed 30 Dec. 2020.

NWS b. Flash Flooding Definition. Retrieved from https://www.weather.gov/phi/FlashFloodingDefinition. Accessed 30 Dec. 2020.

NWS c. NWS JetStream MAX - Addition Köppen Climate Subdivisions. Retrieved from https://www.weather.gov/jetstream/climate_max. Accessed 30 Dec. 2020.

NWS d. Monthly & Annual Precipitation at Central Park. Retrieved from https://www.weather.gov/media/okx/Climate/CentralPark/monthlyannualprecip.pdf. Accessed 30 Dec. 2020. Accessed 30 Dec. 2020.

NWS e. Flash floods and floods...the Awesome Power! Retrieved from https://www.weather.gov/pbz/floods

Open Sewer Atlas NYC. Open Sewer Atlas NYC. Retrieved from https://openseweratlas.tumblr.com/data. Accessed 30 May. 2021.

Petersen, T., Devineni, N., & Sankarasubramanian, A. (2012). Seasonality of monthly runoff over the continental United States: Causality and relations to mean annual and mean monthly distributions of moisture and energy. *Journal of Hydrology*, *468–469*, 139–150. https://doi.org/10.1016/j.jhydrol.2012.08.028

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

Schmitt, T. G., Thomas, M., & Ettrich, N. (2004). Analysis and modeling of flooding in urban drainage systems. *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2004.08.012

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

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

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

700 State of New York. New York State Climate Hazards Profile. Retrieved from
701 https://ap.buffalo.edu/content/dam/ap/PDFs/NYSERDA/New-York-State-Climate-Hazards-
702 Profile.pdf. Accessed 30 Dec. 2020.

703 Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of Royal*
704 *Statistical Society*, *58*(1), 267–288. https://doi.org/10.1093/nq/s3-V.128.490-a

705 United States Census Bureau (2012). Largest Urbanized Areas With Selected Cities and Metro
706 Areas. Retrieved from https://www.census.gov/dataviz/visualizations/026/. Accessed 30
707 Dec. 2020.

708 United States Census Bureau. U.S. Census Bureau QuickFacts: New York city, New York; Bronx
709 County (Bronx Borough), New York; Kings County (Brooklyn Borough), New York; New
710 York County (Manhattan Borough), New York; Queens County (Queens Borough), New
711 York; Richmond County (Staten Isl. Retrieved from
712 https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxborou
713 ghnewyork,kingscountybrooklynboroughnewyork,newyorkcountymanhattanboroughnewyo
714 rk,queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST
715 045219. Accessed 30 Dec. 2020.

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

717 World Meteorological Organization. Development and Implementation of International and
718 Regional Flash Flood Guidance (FFG) and Early Warning Systems. Retrieved from
719 http://www.wmo.int/pages/prog/hwrp/flood/ffgs/documents/FFG_Project_Brif.pdf.
720 Accessed 30 Dec. 2020.

721 Zahura, Faria T., et al. “Training Machine Learning Surrogate Models from a High-Fidelity
722 Physics-Based Model: Application for Real-Time Street-Scale Flood Prediction in an Urban
723 Coastal Community.” *Water Resources Research*, vol. 56, no. 10, 2020,
724 doi:10.1029/2019wr027038.

725 Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal*
726 *of Statistical Software*, *27*(8).

727

728

729

Tables & Figures

Table 1. The percentage of zip codes with the significant predictor for each borough and NYC as total, per category, and the BT Ratio.

	Percent of Zip Codes with Category as Predictor				BT Ratio			
Borough	PRCP	BU	CB	MO	PRCP	BU	CB	MO
All	81	41	47	21				
Queens	93	49	42	29	1.1	0.83	0.79	0.74
Staten Island	100	50	75	17	1.1	0.56	0.65	0.39
Brooklyn	76	46	54	11	1.1	1.5	0.7	0.48
Bronx	83	38	33	17	0.81	0.78	1.8	1.5
Manhattan	60	26	48	21	0.71	1.5	1.4	1.6

Table 2. The average pseudo-R^2, number of zip codes with pseudo-R^2 values within stated intervals for all NYC and each borough.

		Number of Zip Codes				
Borough		(0.43, 0.20]	(0.20, 0.14]	(0.14, 0.09]	(0.09, 0.1	(No Predictors)
All	0.14	37	40	37	51	9
Queens	0.15	17	13	13	15	1
Staten Island	0.22	5	6	1	0	0
Brooklyn	0.16	14	9	7	6	1
Bronx	0.11	0	8	6	9	1
Manhattan	0.08	1	4	10	21	6

Table A: Predictors of significance of NYC by nbGLM

Predictor	p-Value
CB	9.76E-10
BU	7.58E-14
Elevation	5.34E-02

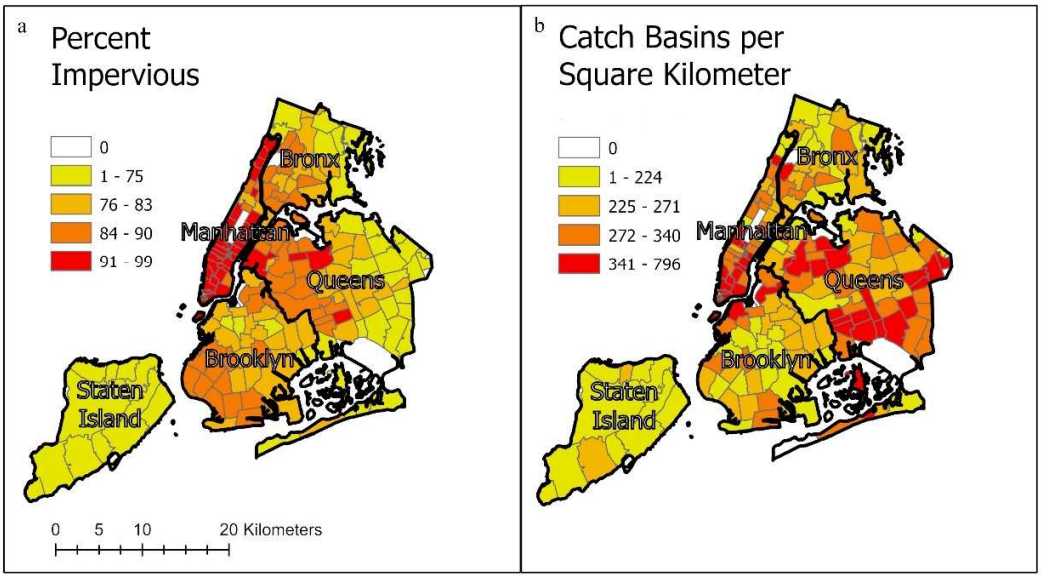
a
Percent Impervious
0
1 - 75
76 - 83
84 - 90
91 - 99
Bronx
Manhattan
Queens
Brooklyn
Staten Island
0 5 10 20 Kilometers
b
Catch Basins per Square Kilometer
0
1 - 224
225 - 271
272 - 340
341 - 796
Bronx
Manhattan
Queens
Brooklyn
Staten Island

Figure 1

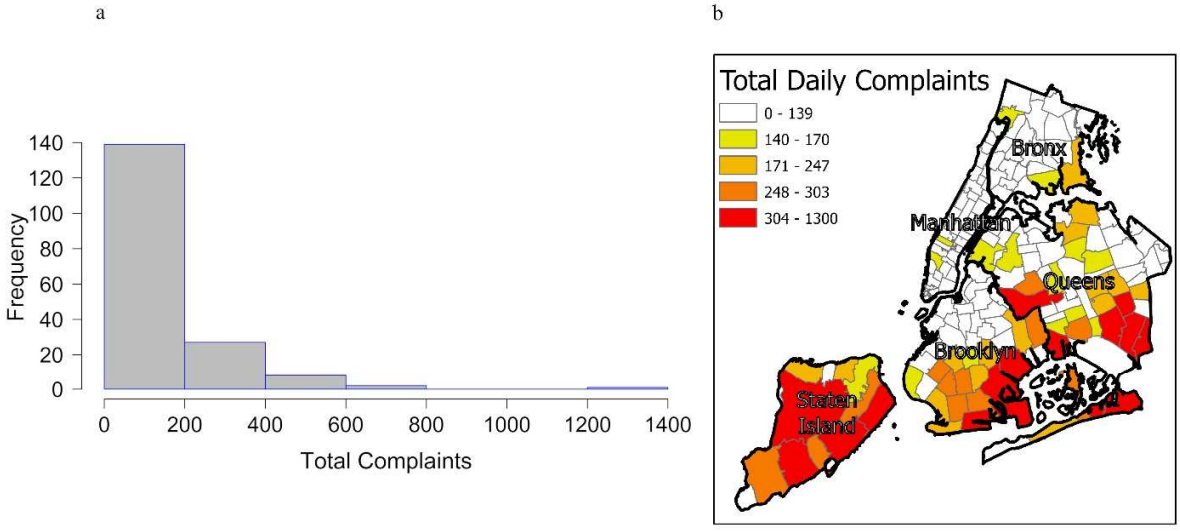
a
b
Frequency
140
120
100
80
60
40
20
0
0
200
400
600
800
1000
1200
1400
Total Complaints
Total Daily Complaints
0 - 139
140 - 170
171 - 247
248 - 303
304 - 1300
Bronx
Manhattan
Queens
Brooklyn
Staten Island

Figure 2

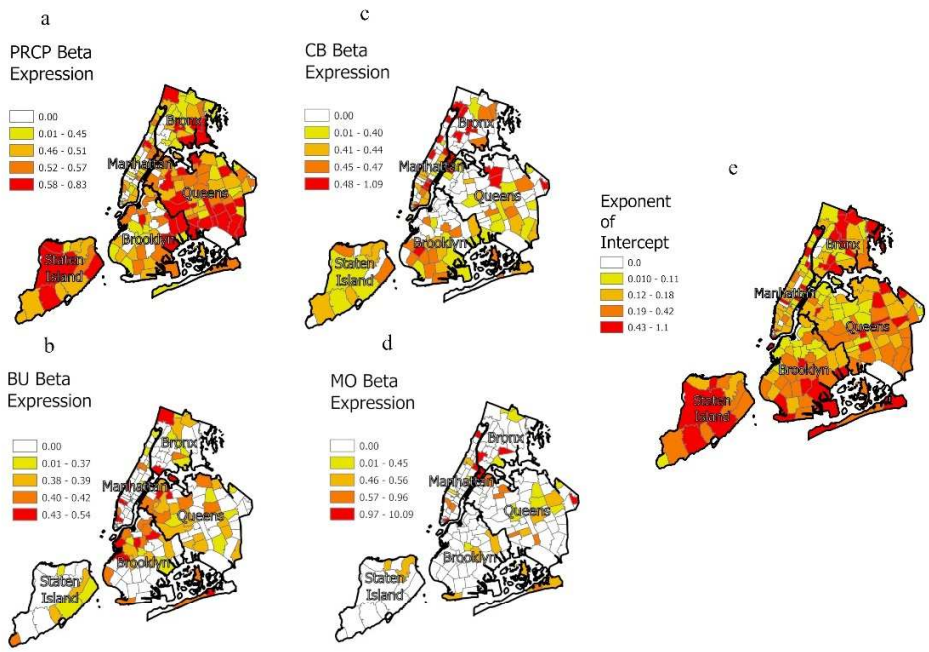
a
PRCP Beta Expression
0.00
0.01 - 0.45
0.46 - 0.51
0.52 - 0.57
0.58 - 0.83
c
CB Beta Expression
0.00
0.01 - 0.40
0.41 - 0.44
0.45 - 0.47
0.48 - 1.09
e
Exponent of Intercept
0.0
0.010 - 0.11
0.12 - 0.18
0.19 - 0.42
0.43 - 1.1
b
BU Beta Expression
0.00
0.01 - 0.37
0.38 - 0.39
0.40 - 0.42
0.43 - 0.54
d
MO Beta Expression
0.00
0.01 - 0.45
0.46 - 0.56
0.57 - 0.96
0.97 - 10.09
Bronx
Manhattan
Queens
Brooklyn
Staten Island

Figure 3

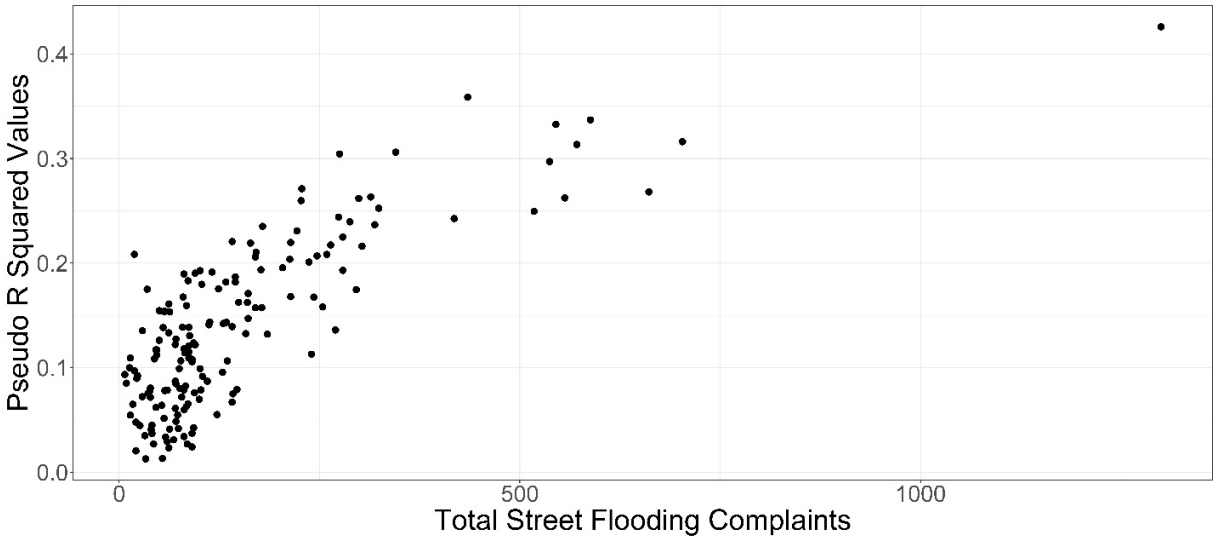

Figure 4

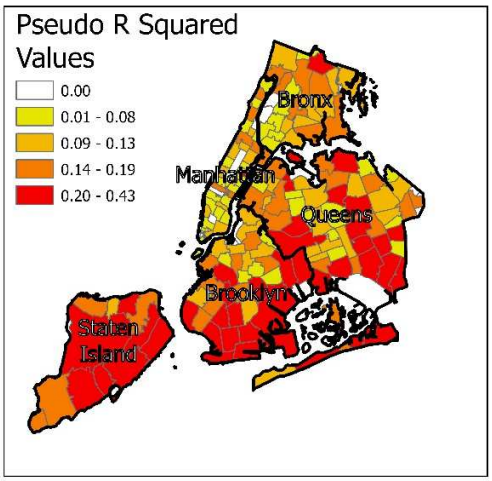

Figure 5

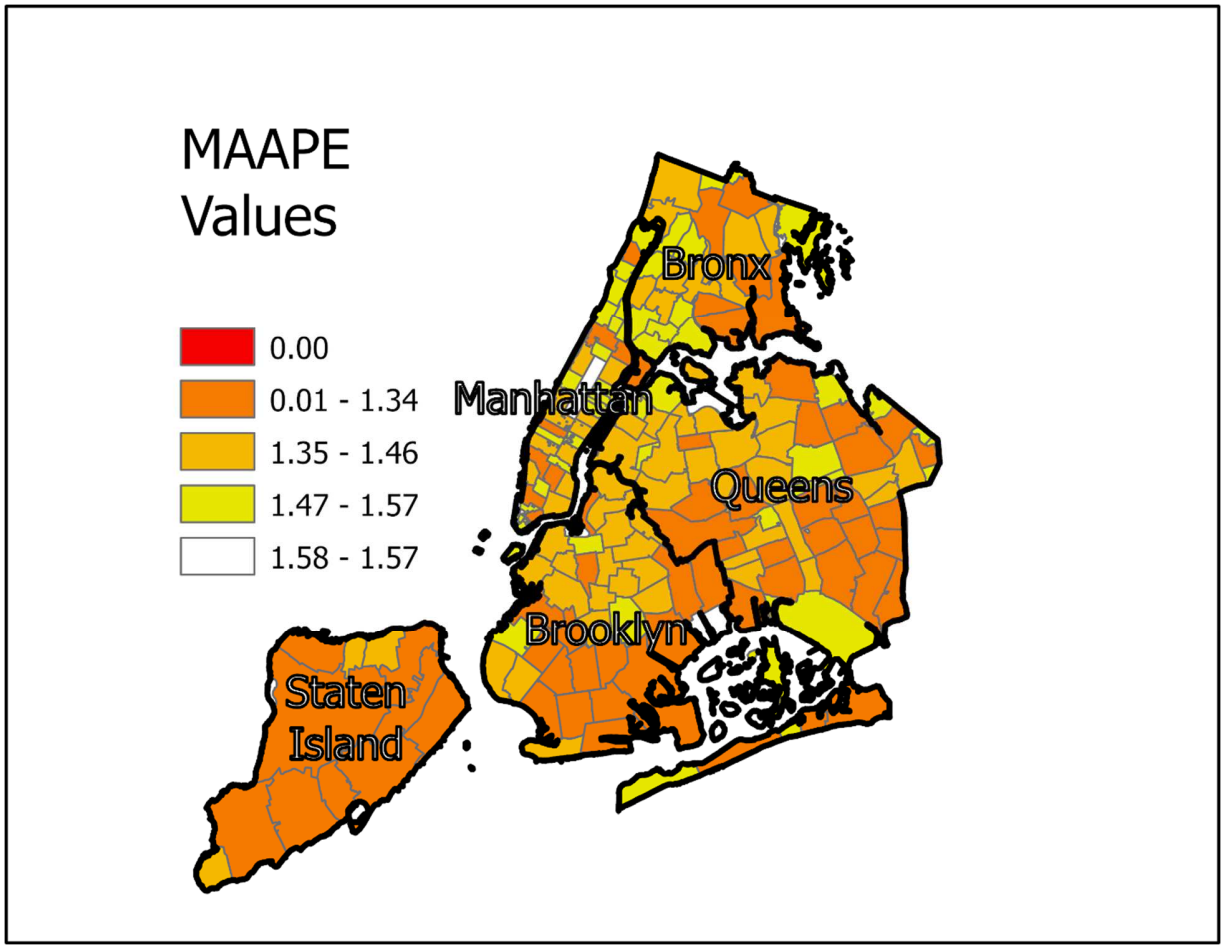
MAAPE
Values
0.00
0.01 - 1.34
1.35 - 1.46
1.47 - 1.57
1.58 - 1.57
Bronx
Manhattan
Queens
Brooklyn
Staten
Island

Figure 6