Regression Analysis of the Variation in Rainfall Derived Inflow and Infiltration

Li Zhang, Fang Cheng, Gregory Barden, Hunter Kelly, Timothy Fallara and Edward Burgess

Rainfall derived inflow and infiltration (RDII) into sanitary sewers is often a major factor contributing to sanitary sewer overflow (SSO) and water in basement (WIB) complaints. SSO poses serious problems by contaminating the environment, causing property damage and threatening public health. Control of SSO is therefore a priority of many sewer system agencies throughout the United States and Canada. Sewer collection systems are commonly modeled to facilitate the control of SSO. Understanding the variation in RDII as a function of rainfall conditions is critical to improving the simulation results of sanitary sewer system models.

This chapter presents a statistical analysis of variation in RDII and its relationship to rainfall characteristics and other weather conditions. As part of the City of Columbus’s long term collection system modeling, sewer flow and rainfall data have been collected for a period of nearly ten years, providing a large dataset for analysis. RDII was analyzed using the USEPA’s recently released sanitary sewer overflow analysis and planning (SSOAP) toolbox program to generate the total RDII capture fraction $R$, which is commonly used to model RDII in sanitary sewer systems, as well as the rainfall characteristics. Multivariate regression analysis was performed on these data using Minitab 15 in order to examine the relationship between the total $R$ and rainfall characteristics such as rainfall volume, rainfall duration, peak intensity and antecedent dry days, and other weather factors such as temperature. A generalized linear regression model was used to reveal the possibility of nonlinear relationships.
The results showed significant relationships between total $R$ and rainfall volume, antecedent dry days, peak intensity and temperature. Seasonal or monthly unit hydrograph (RTK) parameters for RDII are typically used in SWMM 5 modeling for long term continuous simulation of RDII in sanitary sewers. Integration of an RDII regression model into a SWMM 5 model could enable storm-specific RTK values to be used, which might improve the continuous simulation of sanitary sewer systems.

13.1 Introduction

Sanitary sewer collection systems collect and transport wastewater from its sources to wastewater treatment plants. Severe weather, improper system operation and maintenance, and vandalism may cause unintentional discharges of raw sewage from municipal sanitary sewers. These unintentional discharges, or sanitary sewer overflows, occur in almost every system, and there are at least 40 000 SSOs each year (USEPA, 2010). The untreated sewage may cause serious water quality problems, back up into basements, damage property and threaten public health. USEPA administers the national pollutant discharge elimination system permitting program to regulate SSO.

Sewer pipeline stoppages, collapses, aging and deterioration result in excessive inflow and infiltration into a sanitary sewer system and this adversely affects operation of the entire sewerage system (Lai, 2008). Inflow and infiltration (I/I) may greatly increase flows in the sanitary sewer system and cause unnecessary burdens on the treatment plants. SSO occurs when the flows exceed the capacity of a sewer, and result in sewer surcharges. SSO is prevalent during and immediately after wet weather when flows are greatly increased due to excess I/I. That portion of rainfall which gets into a sewer system in the form of inflow or infiltration (RDII) is the major source of excess flow in the sanitary sewer system. In addition to RDII, other contributing factors to SSO include sewer blockage from root intrusions, grease build-up, sedimentation and debris (Lai, 2008). Significant RDII results in increased operating costs and requires higher system capacities than would otherwise be needed. The ability to accurately estimate RDII will help to create effective SSO control plans.

The response of a sanitary sewer system to rainfall is complex. RDII is influenced by rainfall characteristics such as volume, intensity and duration, and its temporal and spatial variation; by antecedent conditions; and by physical conditions including depth to ground water, depth to bedrock, land slope, number and size of sewer system defects, type of storm drainage system, soil characteristics, and type of sewer backfill (Vallabhaneni et al., 2007).
Wastewater flows have two components, one associated with flows during dry weather periods and the other associated with extraneous flows from wet weather events. The component related to dry weather includes base wastewater flow (BWF) and ground water infiltration (GWI). BWF is the residential, commercial, institutional and industrial flow discharged into a sanitary sewer system. GWI is the infiltration of ground water that enters the sanitary sewer system through leaking pipes, pipe joints and manhole walls (Vallabhaneni, 2007). The component related to wet weather is RDII. RDII starts from zero, increases during rainfall and reaches its peak after the rainfall reaches its highest intensity, and then decrease to zero sometime after the end of the rainfall. Figure 13.1 shows the components of wet weather wastewater flow, and RDII curve as a response of rainfall curve. RDII is critical in designing sanitary sewer systems. A reliable estimation of the current and future BWF and RDII insure that sewers have adequate conveyance capacities for the determined design horizon (Lai, 2008).

![Figure 13.1 Three components of sanitary wastewater flow (from Vallabhaneni et al, 2007).](image)

Various methods have been developed to quantify RDII. USEPA has provided a report for a companion literature review of RDII prediction methodologies (Lai, 2008). Methods include the constant unit rate method, the percentage of rainfall volume (R-value) method, the percentage of stream flow method, the probabilistic method, and the rainfall–flow regression method (Vallabhaneni, 2007). Hydraulic modeling software uses one or
more of the prediction methods. SWMM incorporates the synthetic unit hydrograph method; it quantifies and characterizes RDII using three families of unit hydrographs (the RTK method). This is probably the most popular synthetic unit hydrograph method. $R$ is the fraction of rainfall volume entering the sewer system as RDII during and immediately after the rainfall event, $T$ is the time to peak, and $K$ is the ratio of the time of recession to the time to peak (Vallabhaneni, 2007). Figure 13.2 shows the representation of RDII by three sets of unit hydrographs using the RTK method. The first triangle represents the most rapidly responding inflow component; the second includes both rainfall derived inflow and infiltration; and the third includes infiltration that may continue after the storm event ends. The sum of the three $R$ values for the three unit hydrographs ($R_1$, $R_2$ and $R_3$) is the total $R$ value for the rainfall event.

![SUMMATION OF THREE UNIT HYDROGRAPHS](image)

Figure 13.2 RDII characterized by three sets of unit hydrographs (from Vallabhaneni et al, 2007).

### 13.2 Background

This study is supported by the City of Columbus sewer system capacity model (SSCM) update project, to update the model from 2000 to 2008 conditions. One goal is to reduce I/I and eliminate SSO. Over 200 flow meters in the sewer system monitor flow for 16 months; about 30 rain gauges provide rainfall information. Figure 13.3 shows the submodel basins, and the coverage of flow meters and rain gauges in the system. The flow meter data and rain gauge data were used for wet weather flow calibration, including RDII quantification. Some of the meters and the rain gauges have been collecting data for over ten years. This long term data for flow and rainfall makes it possible to analyze the variation in RDII in Columbus.
Regression Analysis of the Variation in Rainfall Derived Inflow and ...

Figure 13.3 Model basins, flow meters and rain gauges in Columbus.

The USEPA SSOAP toolbox provides a statistical analysis function to calculate RDII parameters including rainfall volume, rainfall duration, RDII duration, peak rainfall intensity, rainfall volume for 7 d prior to the event, rainfall volume for 14 d prior to the event, and rainfall volume for 21 d prior to the event, as well as three sets of $R$, $T$ and $K$ for fast, medium and slow responses of RDII to rainfall. Figure 13.4 and Figure 13.5 show the function of SSOAP toolbox for RTK analysis and statistical analysis.

Figure 13.4 RDII analysis function in SSOAP toolbox.
The rainfall–flow regression method is promising if there are long term flow rate data and extensive rainfall data in terms of both time and space to develop regression models to reflect seasonal influences for dry and wet antecedent and groundwater conditions (Vallabhaneni, 2007). Temperature can represent climate change and the vegetation coverage, and long term temperature data could be obtained from a public weather website. In this study, long term temperature data were used to represent the seasonal influences for dry and wet antecedent and groundwater conditions. The historic temperature data at Port Columbus International Airport (KCMH) weather station were obtained from www.wunderground.com, the Weather Underground website.

13.3 Regression Analysis

13.3.1 Generalized Linear Model

Regression analysis uses a statistical methodology to predict a response variable from the other predictor variables, utilizing the relationship between them. A linear regression model is defined as:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} + \epsilon_i \]  

(13.1)

where:

- \( Y_i \) = response variable,
Regression Analysis of the Variation in Rainfall Derived Inflow and...

\[X_{i1}, X_{i2}, \ldots, X_{ip-1} = \text{values of the } p-1 \text{ predictors for the } i\text{th case},\]

\[\beta_0 = \text{constant or intercept},\]

\[\beta_j = \text{regression coefficient for the corresponding } j\text{th predictor } X_j, \text{ and}\]

\[\varepsilon_i = \text{error term reflected in the residuals for the } i\text{th case}.\]

The regression function for the model is defined as:

\[E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}\]  \hspace{1cm} (13.2)

where:

\[E\{Y\} = \text{mean response (Kutner et al., 2004).}\]

The accuracy of a given regression equation could be assessed by several statistical parameters. Analysis of variance (ANOVA) provides information about the sources in the variation of the data. The coefficient of multiple determination \(R^2\) measures the percentage of the variance in the response variable \(Y\) explained by the predictors, and is defined as:

\[R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}\]  \hspace{1cm} (13.3)

where:

\[SSR = \text{regression sum of squares } \sum(Y-\hat{Y}),\]

\[SST = \text{total sum of squares } \sum(Y-\bar{Y}),\]

\[SSE = \text{error sum of squares } \sum(\hat{Y}-Y),\]

\[\hat{Y} = \text{regression prediction for the } i\text{th case, and}\]

\[\bar{Y} = \text{mean of the response values.}\]

\(R^2\) is increased and never decreased by adding more predictors, because \(SSE\) can never become larger with more predictors and \(SST\) is always the same for a given set of responses. The adjusted coefficient of multiple determination \(R_a^2\) is a modified measure by adjusting for the number of predictors in the model, described by the equation:

\[R_a^2 = 1 - \frac{(n-1) \frac{SSE}{SST}}{n - p}\]  \hspace{1cm} (13.4)

where:

\[n = \text{sample size, and}\]

\[p = \text{number of terms in the model, including the predictors and the constant value.}\]

The \(F\)-hypothesis test for the coefficients on all the predictors is used to test the significance of the regression model as a whole. The test statistic is:

\[F_{df_1, df_2} = \frac{SSE/(p-1)}{SSE/(n-p)}\]  \hspace{1cm} (13.5)
The greater the value of $F$, the more significant the regression equation will be. For the multiple regression equation to be statistically significant, the value of the computed $F$ must be greater than a critical value $F_{cr}$. The value of $F_{cr}$ is taken from the $F$ distribution table and depends on the number of degrees of freedom and a significance level $\alpha$ (usually $\alpha = 0.05$).

When the response and the predictors have a nonlinear relationship, transformations for linearizing a nonlinear regression relation may be used. All of the regression models, including linear and nonlinear, belong to the generalized linear models family. In this study, the response is the total $R$, as a percentage ($0 < R < 1$). The predictors may be rainfall volume, rainfall duration, peak rainfall intensity, rainfall volume for 7 d prior to the event, rainfall volume for 14 d prior to the event, rainfall volume for 21 d prior to the event, and the average of daily mean temperature for 10 d prior to the event. The relationship of the response to the variable would not be linear since in practice there are lower and upper limits for RDII. For a specific flow meter basin, total $R$ may belong to $(0, \theta)$, with an upper limit value $\theta$ where $0 < \theta < 1$. A transformation is applied on total $R$:

$$Y' = \ln\left(\frac{Y}{(\theta - Y)}\right) \quad (13.6)$$

where:

- $Y$ = total $R$ for RDII,
- $Y'$ = transformed total $R$, and
- $\theta$ = upper limit of total $R$ ($0 < \theta < 1$).

For the flow meters in Columbus, the upper limit for total $R$ is about 30%.

### 13.3.2 Case Study

Rain gauge data and flow data were examined for consistent data quality in the long term. Long term rainfall data at a nearby rain gauge was used to represent the rainfall for the flow meter basin due to the lack of long term radar rainfall data. Small areas of flow meter basins are preferred to limit the error induced by spatial variation of rainfall. The flow meter basin Columbus_Teteridge was selected for a case study of multivariate regression analysis of RDII variation in Columbus.

Figure 13.6 shows the Columbus_Teteridge basin with the flow meter Columbus_Teteridge and the rain gauge DIAD4450 used for that basin. This basin has a contributory area of 3.55 km$^2$. The average age of the pipes in the area, built around 1961, is 50 y. The major land use is residential. It is a leaky area located within Clinton #3 subbasin in Columbus. The long term flow data used are for the years 2002–2004 and 2007–2009. Flow rate data
and rainfall data with 15 min time steps were put into the SSOAP toolbox and rainfall and RDII events were defined. The statistical analysis function in SSOAP toolbox was used to generate the parameters for regression analysis. In total 65 storms were analyzed, 38 storms in the growth season (May–October) and 27 in the dormant season (November–April). Figure 13.7 shows the total $R$ for the 65 storms with each storm represented by a dot.

![Figure 13.6 Columbus_Teteridge flow meter basin.](image)

![Dotplot of R](image)

Figure 13.7 Total $R$ for 65 storms at Columbus_Teteridge.

### 13.4 Results and Discussion

Multivariate regression analysis was performed on the 56 storms for the Columbus_Teteridge flow meter basin to examine the relationship between the total $R$ and rainfall characteristics such as rainfall volume, rainfall duration, peak intensity and rainfall in previous days, and temperature.
Analysis was performed for the combined seasons with whole data set of all the storms, and also for both the growth season and winter season separately. Best regression models were selected for each of them based on the regression model criteria described above.

The equation of the regression model for combined seasons using all of the 65 storms is:

\[ TrsR = 1.54 + 0.576 \text{Vol} + 0.316 \text{Rn21} – 1.91 \text{ExpTemp} \]  \hspace{1cm} (13.7)

where:

\[ TrsR \] = transformed total R using Equation 13.6,
\[ Vol \] = rainfall event volume,
\[ Rn21 \] = 21 d rainfall prior to the event, and
\[ ExpTemp \] = exponential transformation of temperature.

The temperature data is the average of daily mean temperatures in 10 days prior to the storm event including the day when the storm started.

All the three predictors are significant with \( p \approx 0 \). The adjusted \( R^2 \) is 40.0, meaning that 40% of the variation of RDII could be explained by the predictors. The \( p \)-value for the \( F \)-test for the regression is close to 0 indicating the significance of the whole regression model. Figure 13.8 shows that the model residuals satisfy the requirement of linear regression that the residuals have normal distribution with zero mean and constant variation.

Figure 13.8 Residual plots for the combined model at Columbus_Teteridge.
After separating the storms into growth and dormant seasons, the regression analysis had two separate regression models for the two seasons.

The equation of the regression model using the 38 storms in the growth season is:

\[
T_{rsR} = 3.62 + 0.387 Vol + 0.472 R_{n21} - 2.86 \text{ExpTemp} \quad (13.8)
\]

The equation of the regression model using the 27 storms in the dormant season is:

\[
T_{rsR} = 1.14 + 2.46 Vol - 5.75 PI - 1.49 \text{ExpTemp} \quad (13.9)
\]

Figures 13.10 and 13.11 show that the model residuals satisfy the requirement of linear regression that the residuals have normal distribution with mean = 0 and constant variation. Both the regression models are significant according to the \(F\)-test with \(p \approx 0\). For both the growth season and dormant season models, all the three predictors are significant with \(p \approx 0\). The predictor \(R_{n21}\) for 21-day rainfall prior to the event was not a predictor in the dormant season model, but \(PI\) for peak rainfall intensity was added as a significant predictor. The adjusted \(R^2\) is slightly improved at 42.8 for the growth season model, and greatly improved at 66.2 for the dormant season model. A comparison of Figure 13.9 with Figure 13.12 shows that after separating the regression models into growth season and dormant season, the fitted total \(R\) by the models became closer to the the observed total \(R\) especially for the rainfall events with greater total \(R\).
Regression Analysis of the Variation in Rainfall Derived Inflow and ...

Figure 13.10 Residual plot for growth season model at Columbus_Teteridge.

Figure 13.11 Residual plot for dormant season model at Columbus_Teteridge.
All three models show that rainfall volume is a significant predictor of the variation of RDII. Temperature representing the seasonal change of vegetation coverage and climate is also a significant predictor of the variation of RDII. The coefficient of temperature in the growth season model is greater than it is in the dormant season model, which indicates that vegetation coverage affects the variation of RDII more than it does in the dormant season, which is reasonable since vegetation will absorb and hold more rainfall in the growth season.

Peak intensity is a significant predictor in the dormant season model but not in the growth season model. Rainfall prior to the event is a significant predictor in the dormant season model but not in the growth season model. Vegetation coverage is greatly reduced in the dormant season, so the antecedent condition has less and peak rainfall intensity has more impact on RDII.

The accuracy of the regression model may be decreased by the errors introduced with the rainfall data and the flow data. The rain gauge data may not accurately represent precipitation in the flow meter basin due to spatial variation of rainfall. Using radar rainfall data may be more accurate. Due to the lack of long term radar rainfall data, a case study was performed using short term rainfall for 2008 to 2009 in a flow meter basin of SC-MI-12 meter with a small contributary area of 1.85 km². It showed that the adjusted $R^2$ could be improved from 51.3 to 78.0 for combined models using all the storms.
In this study, the RTK analysis in the SSOAP toolbox does not include the initial abstraction parameters, which indicates that the antecedent condition information is implicitly included in the RTK values. Variation of RDII is characterized by RTK only, since no initial abstraction counted.

Overall, the significance of the regression model indicated that total $R$ varies with different values of predictors, which means that RDII varies with the varying rainfall condition, such as volume and peak intensity, as well as the varying seasonal condition such as climate and vegetation coverage. The regression model itself including the equation structure and coefficients represents the physical system condition of the flow meter basin such as pipeline aging and deterioration.

13.5 Conclusion and Recommendation

From the regression analysis, we conclude that a significant relationship exists between total $R$ and the parameters rainfall volume, temperature, and rainfall in previous days prior to the event. Growth and dormant seasons have different regression relationships. Separating the regression analysis into growth season and dormant season could improve the regression analysis by explaining more variation in RDII. In the growth season, more than 40% of the variation of RDII could be explained by rainfall volume, temperature, and previous rainfall prior to the event. In the dormant season, more than 60% of the variation of RDII could be explained by rainfall volume, temperature, and peak rainfall intensity. Prediction of total $R$ using the regression model may improve the continuous simulation by using storm specific $R$.

The regression model may also be improved by testing on more flow metered basins in Columbus and by using spatial rainfall data such as radar data for the study area.

References


USEPA. (2010). National Pollutant Discharge Elimination System (NPDES), Sanitary Sewer Overflows.