NEW YORK STATE
WATER RESOURCES INSTITUTE
Department of Biological and Environmental Engineering

230 Riley-Robb Hall, Cornell University
Ithaca, NY 14853-5701
http://wri.cals.cornell.edu

Tel: (607) 254-7163
Fax: (607) 255-4080
Email: nyswri@cornell.edu

Analyzing the discrepancy between return period stream flows using the TR - 55 Method and USGS recorded stream discharges

Karin Teuffer, M.Eng Civil and Environmental Engineering, kt428@cornell.edu

Small culvert in need of maintenance and inspection

Abstract

Stream flows calculated with the TR-55 method and obtained from USGS empirical data were compared for twenty stream gauges and their corresponding watersheds in New York State. The overall differences between the distributions of the two methods were measured using the Kolmogorov-Smirnov statistic, and 1, 10, and 100 year return period percent differences in flows. The three variables regressed against were average curve number, latitude and drainage area. Multiple linear regression, and lasso and ridge regression showed that none of the predictor variables had significant influence on the difference between modeled and the measured values. Area and latitude have higher correlations with the raw flows for the three return period storms than curve number when analyzed individually. This was not seen in multiple regression and may be representative of a curve number influence on the discrepancy between model and empirical data.

Three Summary Points of Interest

- **The small quantity of watersheds analyzed is an important factor to consider in this analysis. A greater number of locations, even further diversified in terms of latitude and drainage area, could yield more credible results.**

- **The limiting obstacle in this case was processing time in ArcMap due to the size of the 10 meter DEM.**

- **The results show that neither curve number, nor latitude or drainage area, heavily affects the discrepancy between the TR-55 model and the empirical stream flow values recorded by the USGS**.

Keywords
USGS stream gauges, Multivariate analysis, Hydrograph model comparison

Introduction

Comparing a model to empirical data is a form of attaining credibility as well as determining the parameters that can reflect the dynamic nature of real world scenarios. When using environmental data, it is often difficult to discern the individual effects of the predictors because they tend to be highly correlated. Thus, statistical methods can help determine which variables have the greatest influence over the model's output and if the inputs match the overall effects of environmental variables seen in nature. Regression analysis is one method for gauging the discrepancy, and is used in this report to analyze which out of three tested variables contribute the most to distribution differences of stream flow data obtained through two methods. The three predictors are curve number, latitude, and drainage area. The dependent variable is either the Kolmigorov-Smirnov statistic, or 1, 10, and 100 year return period flow percent differences. The two datasets being compared are one calculated using the TR-55 and triangular hydrograph method and an empirical one obtained from the USGS recorded instantaneous peak flows.

The models compared are relevant to current studies conducted on culverts within New York State. The capacity of culverts compared to the flows they must convey, is an important factor to be studied due to their purpose in mitigating flooding and other disruptions, as well as their function as ecological connections. Culverts that may have the correct capacity at present, might nevertheless be on the verge of failing and thus be vulnerable to the increasing intensity of storms seen to arise from climate change.

Work by Marjerison *et al.* (unpublished) suggests that there are areas in the Lower Hudson River watershed whose culverts are not able to convey peak flows for return period storms of 5, 10 or 100 years. A suitability threshold was placed at a capacity for flows corresponding to a 5-year storm. The results show that 62 percent of the culverts had at least this capacity. This was an overall number, however, and the study delved deeper into what caused differences between suitability percentages by watershed. They found that median income and state road ownership were positively correlated with suitability.

Likewise, the New York State Water Resources Institute has been working on developing and enhancing a computer model using Python and ArcMap to determine the capacity of culverts within the state. Field data on culvert dimensions, type, maintenance state, location, ownership and many other characteristics are collected first, and then used to calculate capacity. The model also compares the results to precipitation curves, for 24-hour duration storms for various return periods, to determine the maximum intensity that each culvert can bear. This is where the model evaluation step commences. To ensure that the method for capacity extraction is accurate, the results were compared against StreamStats values for the same return periods. This yielded a sense that the WRI model mostly overestimates the area of the watersheds and likewise the runoff that each culvert receives. The model parameters modified were the percent of storage that is quantified as the initial abstraction. Decreasing it from 0.2 to 0.03 increased the ratio proximity to 1 but still showed a discrepancy. There were doubts as to whether the interpolation of the StreamStats method was influencing results and whether there were other important factors apart from initial abstraction. Hence, the goal of this study was to evaluate the WRI model with empirical data recorded by the USGS at actual stream gauges. The model steps were completed manually, rather than through Python - ArcGIS modeling tools, and were compared to USGS values.

Results & Discussion

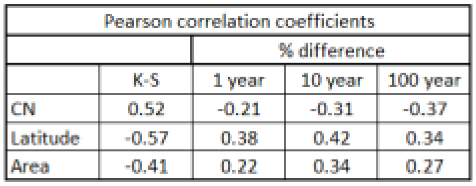
The Pearson correlation coefficients were calculated between all predictor variables, and between all predictor variables and dependent variables. The variance inflation factor (VIF) values were also extracted from regression of all the predictor variables against each other to determine the magnitude of the multicollinearity. Usually a VIF value equal to or greater than 10 is considered a problem for model development because it signifies that the variance of the predictor estimates are increased by an order of magnitude. That is not the case here. Table 1 shows that the VIFs are relatively insignificant and supports the finding that most of the highest multicollinearity effects occur with latitude. Table 2 shows that the curve numbers and both latitude and drainage area are negatively correlated, while latitude and drainage area are positively correlated with each other. Table 3 shows that the sign of the correlation coefficients differs significantly between the Kolmigorov-Smirnov statistic (K-S variable) and the percent difference variables for all return periods. Latitude and area are positively correlated with the percent differences while they are both negatively correlated with K-S values. In addition, latitude is again the greatest in magnitude in all cases.

	VIF
CN	1.776
Latitude	2.282
Area	1.458

Table 1

Pearson correlation coefficients			
	CN	Latitude	Area
CN	1.00	-0.66	-0.34
Latitude	-0.66	1.00	0.56
Area	-0.34	0.56	1.00

Table 2



Pearson correlation coefficients				
		% difference		
	K-S	1 year	10 year	100 year
CN	0.52	-0.21	-0.31	-0.37
Latitude	-0.57	0.38	0.42	0.34
Area	-0.41	0.22	0.34	0.27

Table 3

The coefficient estimates from linear multiple regression further support the correlation values. The signs of the estimates in Table 4 reflect those of Table 3. The fact that the variables were centered and scaled before analysis helps reduce the absolute standard error. The p values for all coefficient estimates are all greater than 0.25. P values are usually not the best measure of each predictor because multicollinearity affects how much each variable contributes to the total variance of the dependent variable. When two variables are correlated, their contribution is divided and reflected in increased p values. However, in this case the multicollinearity was shown to be low.

One thing to note is that as the return period used for percent difference increases, the effect of latitude seems to decrease. The correlation is lower and the p value is higher. These 100-year return period values were also those that were extrapolated the most from fitting the data to the Log Pearson III distribution.

KS coefficient estimates			
	Estimate	Std. Error	p
(Intercept)	0.00	0.19	1.00
CN	0.27	0.26	0.33
Latitude	-0.31	0.30	0.31
Area	-0.15	0.24	0.55
1 year % difference			
	Estimate	Std. Error	p
(Intercept)	0.00	0.23	1.00
CN	0.07	0.31	0.82
Latitude	0.42	0.35	0.25
Area	0.01	0.28	0.98
10 year % difference			
	Estimate	Std. Error	p
(Intercept)	0.00	0.22	1.00
CN	-0.08	0.30	0.80
Latitude	0.27	0.34	0.43
Area	0.16	0.27	0.56
100 year % difference			
	Estimate	Std. Error	p vlaue
(Intercept)	0.00	0.22	1.00
CN	-0.26	0.30	0.41
Latitude	0.11	0.35	0.76
Area	0.12	0.28	0.67

Table 4

Table 5 holds the lasso and ridge regression coefficients for the same set of centered and scaled data. These methods are used when multicollinearity increases the variance of the estimators to a point in which interpretation of the results is difficult. With addition of a parameter lambda the magnitude of the covariance matrix values are controlled using penalties on the constraints. When two variables are correlated, the lasso method chooses one and maximizes its coefficient while reducing the other coefficient to zero. Ridge regression splits the contribution of two correlated variables and reduces the magnitude of both coefficients. The K-S case reflects the linear regression coefficients. This is likely due to how the statistic measures a broader discrepancy between both distributions of stream flows. The ridge regression coefficients are very low for the percent difference cases and only slightly reduced, compared to the linear and lasso results, in the K-S case. For lasso regression, the 10 and 100-year percent difference cases suggest that none of the variables significantly contribute to the variance. Only latitude has a role in the one year percent difference case.

Lasso coefficient estimates				
	KS	1 year %	10 year %	100 year %
(Intercept)	0.00	0.00	0.00	0.00
CN	0.25	0.00	0.00	0.00
Latitude	-0.30	0.28	0.00	0.00
Area	-0.13	0.00	0.00	0.00
Ridge coefficient estimates				
	KS	1 year %	10 year %	100 year %
(Intercept)	0.00	0.00	0.00	0.00
CN	0.20	-0.04	-0.07	-0.11
Latitude	-0.21	0.14	0.11	0.09
Area	-0.13	0.05	0.08	0.07

Table 5

The plots of the scaled variables are not easily interpretable. Figure 1 thus presents all twenty points, of the predictor and dependent variables and their individual relationships, using unscaled values. The regression lines are shown in some of them - these are the cases in which the intercepts were not too large to be out of range of the axes as shown. Their near zero slope nevertheless represents what most of the plots are showing, no relationship between the predictors and their dependent variables.

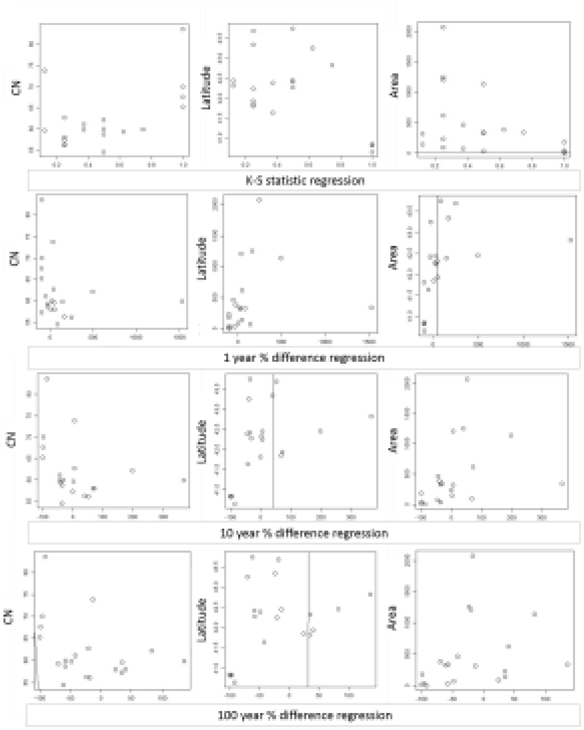
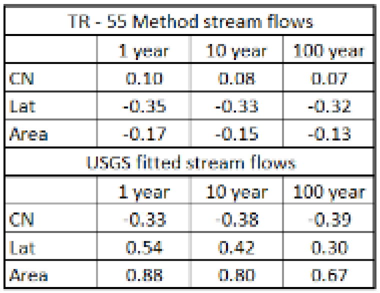


Figure 1

The purpose of this study was to help define or eliminate the predictor variables as contributors to the discrepancy between the TR-55 based WRI model and empirical USGS data. However, to better understand the lack of correlation between the variables and their effect on the model differences it is important to analyse their individual relationships with the raw flows. The correlation between the predictor variables is the same as in previous results. However , Table 6 shows the correlation between the variables and the raw flows for the 1, 10 and 100 year return period storms. The top of the chart represents the values calculated between the variables and the stream flows from the TR-55 method. The bottom portion uses the same variable values but the stream flow variable represents the USGS Log Pearson III fitted values. The results show that the TR-55 method reflects the signs found by using the K-S statistic, and the USGS values reflect the percent difference correlation values discussed above. The main source of inconsistency between the two stream flows would be the curve number which was calculated using the shape file. However, the latitude and area are equivalent. This suggests that the curve number may be the worthwhile variable to explore further.



TR - 55 Method stream flows			
	1 year	10 year	100 year
CN	0.10	0.08	0.07
Lat	-0.35	-0.33	-0.32
Area	-0.17	-0.15	-0.13
USGS fitted stream flows			
	1 year	10 year	100 year
CN	-0.33	-0.38	-0.39
Lat	0.54	0.42	0.30
Area	0.88	0.80	0.67

Table 6

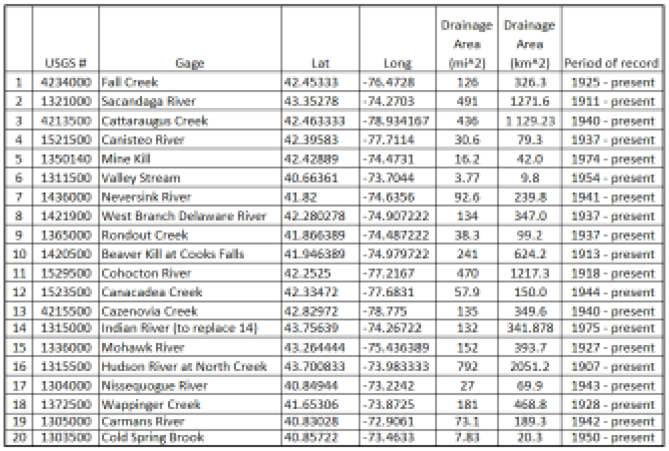
Policy/Management Implications

The preliminary data collection of the culvert characteristics involved field work throughout nearby communities. The recorded information on culvert dimensions, maintenance and current structural state, potential clogging, and other observations were not only inputted into the WRI model discussed above but also shared with the NAACC. Thus, the data helps in both determining the adequacy of culvert capacity and its role in inhibiting or aiding aquatic connectivity for organisms.

The statistical research in this report will hopefully allow for the continuing improvement of the WRI culvert capacity model. Once it has been proven to be the most accurate it can be, it will be able to serve communities by indicating the culverts capable of facilitating flooding. This knowledge may then be used by local governments and organizations to improve the design of, replace, or transfer management of the culverts to the appropriate party .

Methods

To have consistent locations for comparison between the two methods, twenty stream flow gauge sites were chosen from the USGS online database. These points were chosen based on a significant amount of recording history, at least thirty years, varying drainage area size, and dispersion over the whole state. Table 7 lists the locations and their main attributes, and Figure 2 displays them on a map of the state.



	USGS #	Gage	Lat	Long	Drainage Area (mi^2)	Drainage Area (km^2)	Period of record
1	4234000	Fall Creek	42.45333	-76.4728	126	326.3	1925 - present
2	1321000	Sacandaga River	43.35278	-74.2703	491	1271.6	1911 - present
3	4213500	Cattaraugus Creek	42.463333	-78.934167	436	1 129.23	1940 - present
4	1521500	Canisteo River	42.39583	-77.7114	30.6	79.3	1937 - present
5	1350140	Mine Kill	42.42889	-74.4731	16.2	42.0	1974 - present
6	1311500	Valley Stream	40.66361	-73.7044	3.77	9.8	1954 - present
7	1436000	Neversink River	41.82	-74.6356	92.6	239.8	1941 - present
8	1421900	West Branch Delaware River	42.280278	-74.907222	134	347.0	1937 - present
9	1365000	Rondout Creek	41.866389	-74.487222	38.3	99.2	1937 - present
10	1420500	Beaver Kill at Cooks Falls	41.946389	-74.979722	241	624.2	1913 - present
11	1529500	Cohocton River	42.2525	-77.2167	470	1217.3	1918 - present
12	1523500	Canacadea Creek	42.33472	-77.6831	57.9	150.0	1944 - present
13	4215500	Cazenovia Creek	42.82972	-78.775	135	349.6	1940 - present
14	1315000	Indian River (to replace 14)	43.75639	-74.26722	132	341.878	1975 - present
15	1336000	Mohawk River	43.264444	-75.436389	152	393.7	1927 - present
16	1315500	Hudson River at North Creek	43.700833	-73.983333	792	2051.2	1907 - present
17	1304000	Nissequogue River	40.84944	-73.2242	27	69.9	1943 - present
18	1372500	Wappinger Creek	41.65306	-73.8725	181	468.8	1928 - present
19	1305000	Carmans River	40.83028	-72.9061	73.1	189.3	1942 - present
20	1303500	Cold Spring Brook	40.85722	-73.4633	7.83	20.3	1950 - present

Table 7

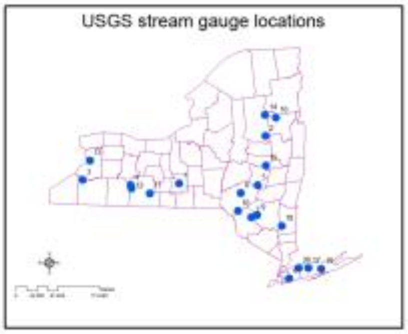


Figure 2

Once the locations were finalized, ArcMap Spatial Analyst-Hydrology tools were used to delineate the watershed, find its longest flow path, and calculate the mean slope. A 10 meter DEM from the NYGIS Clearinghouse was used for the base extraction. A land use-land cover layer from the same site was used to calculate the average curve number for the watershed . From the mean slope (degrees) the average velocity was estimated using Figure 3-1 in the TR-55 (1986) documentation. Storage, initial abstraction (assuming it is 0.2*Storage), and time of concentration were subsequently calculated to yield discharge values, Q, for each return period (NOAA Atlas 14 point precipitation). The USGS stream flow values were calculated by return period using frequency analysis. The annual peak flows were ranked, the exceedance probability, P, was calculated using rank / years on record, and the corresponding return period was found using 1/ P .

The frequency analysis on the USGS peak flows did not yield precise return period stream flows that could be directly compared to those calculated using the TR-55 method. Therefore, before starting the analysis the USGS data was fit to a Log Pearson III distribution using the method of moments.

Regression Analysis
The three predictors found from the ArcGIS data and used as independent variables were average curve number, latitude (of the gauge point), and drainage area . There were four dependent variables on which regression was used: a Kolmigorov-Smirnov statistic and percent differences between the flows of 1, 10 and 100 year return periods for both methods.

Correlation analysis and VIF values were used on all the variables as a preliminary definition of the relationships. Then, linear regression was performed using each dependent variable listed above. Ridge and lasso regressions with the appropriate lambdas (those with the smallest cross validation error) were also preformed.

The best regression method was determined by evaluating which yielded the least variance when \textit{predicting} out of sample values (in this case the test data set consisted of half of the actual data collected).

Student Training

One Master of Engineering student completed the analysis and write up for this study.

Additional final reports related to water resource infrastructure research are available at http://wri.cals.cornell.edu/grants-funding

References

Marjerison, R., Meyer, A., DeGaetano, A., and Walter, M. T. (n.d.). Assessing causes of road culvert conveyance suitability in the Lower Hudson Valley. Ithaca, NY.

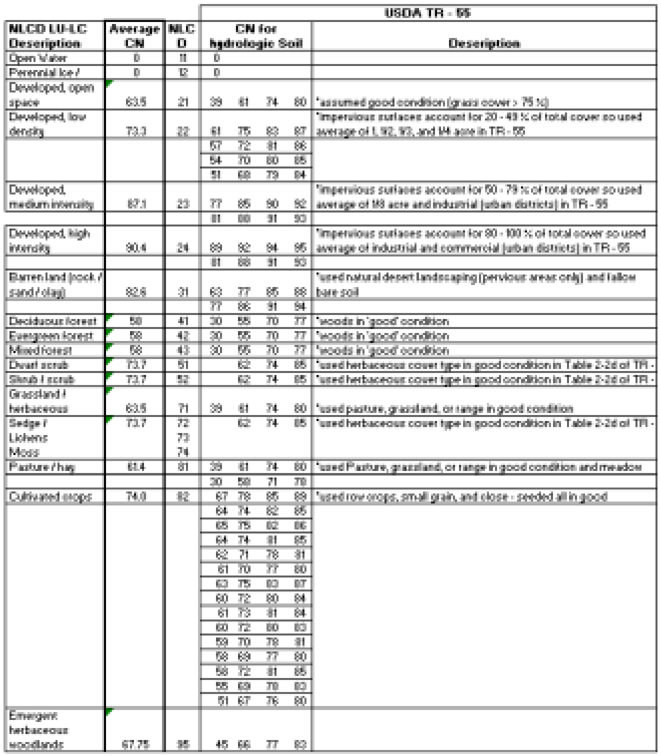
United States Department of Agriculture. (1986). Urban Hydrology for Small Watersheds : TR - 55 (No. Technical Release 55).

Appendices

The average curve numbers for each watershed were calculated by equating the land use-land cover (2006) shapefile metadata descriptions to those in the USDA TR-55 (1986) documentation. The results are shown in the table below.

Note: Only the curve number for the first gauge site (Fall Creek) was found by calibration through SWAT, and not the method described here.

Analyzing the discrepancy between stream flows calculated by two different methods



			USDA TR - 55				
NLCD LU-LC Description	**Average CN**	**NLCD**	**CN for hydrologic Soil**				**Description**
Open Water	0	11	0				
Perennial Ice /	0	12	0				
Developed, open space	63.5	21	39	61	74	80	*assumed good condition (grass cover > 75 %)
Developed, low density	73.3	22	61	75	83	87	*impervious surfaces account for 20 - 49 % of total cover so used average of 1, 1/2, 1/3, and 1/4 acre in TR - 55
			57	72	81	86	
			54	70	80	85	
			51	68	79	84	
Developed, medium intensity	87.1	23	77	85	90	92	*impervious surfaces account for 50 - 79 % of total cover so used average of 1/8 acre and industrial (urban districts) in TR - 55
			81	88	91	93	
Developed, high intensity	90.4	24	89	92	94	95	*impervious surfaces account for 80 - 100 % of total cover so used average of industrial and commercial (urban districts) in TR - 55
			81	88	91	93	
Barren land (rock / sand / clay)	82.6	31	63	77	85	88	*used natural desert landscaping (pervious areas only) and fallow bare soil
			77	86	91	94	
Deciduous forest	58	41	30	55	70	77	*woods in 'good' condition
Evergreen forest	58	42	30	55	70	77	*woods in 'good' condition
Mixed forest	58	43	30	55	70	77	*woods in 'good' condition
Dwarf scrub	73.7	51		62	74	85	*used herbaceous cover type in good condition in Table 2-2d of TR -
Shrub / scrub	73.7	52		62	74	85	*used herbaceous cover type in good condition in Table 2-2d of TR -
Grassland / herbaceous	63.5	71	39	61	74	80	*used pasture, grassland, or range in good condition
Sedge / Lichens Moss	73.7	72 73 74		62	74	85	*used herbaceous cover type in good condition in Table 2-2d of TR -
Pasture / hay	61.4	81	39	61	74	80	*used Pasture, grassland, or range in good condition and meadow
			30	58	71	78	
Cultivated crops	74.0	82	67	78	85	89	*used row crops, small grain, and close - seeded all in good
			64	74	82	85	
			65	75	82	86	
			64	74	81	85	
			62	71	78	81	
			61	70	77	80	
			63	75	83	87	
			60	72	80	84	
			61	73	81	84	
			60	72	80	83	
			59	70	78	81	
			58	69	77	80	
			58	72	81	85	
			55	69	78	83	
			51	67	76	80	
Emergent herbaceous woodlands	67.75	95	45	66	77	83	

This report was prepared for the New York State Water Resources Institute (WRI) and the Hudson River Estuary program of the New York State Department of Environmental Conservation, with support from the NYS Environmental Protection Fund