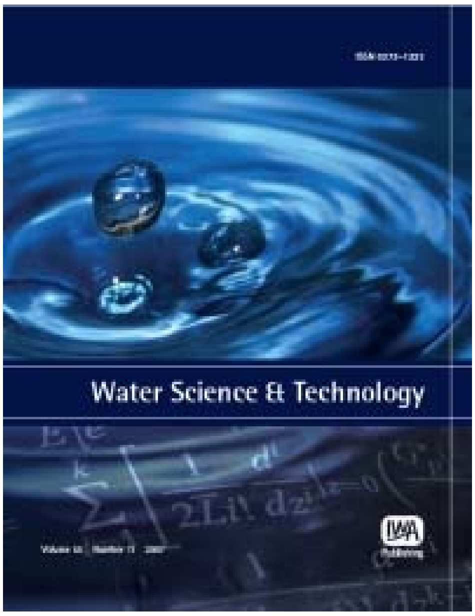
Water Science & Technology

The evaluation of rainfall influence on combined sewer overflows characteristics: the Berlin case study

S. Sandoval, A. Torres, E. Pawlowsky-Reusing, M. Riechel and N. Caradot

ABSTRACT

The present study aims to explore the relationship between rainfall variables and water quality/quantity characteristics of combined sewer overflows (CSOs), by the use of multivariate statistical methods and online measurements at a principal CSO outlet in Berlin (Germany). Canonical correlation results showed that the maximum and average rainfall intensities are the most influential variables to describe CSO water quantity and pollutant loads whereas the duration of the rainfall event and the rain depth seem to be the most influential variables to describe CSO pollutant concentrations. The analysis of partial least squares (PLS) regression models confirms the findings of the canonical correlation and highlights three main influences of rainfall on CSO characteristics: (i) CSO water quantity characteristics are mainly influenced by the maximal rainfall intensities, (ii) CSO pollutant concentrations were found to be mostly associated with duration of the rainfall and (iii) pollutant loads seemed to be principally influenced by dry weather duration before the rainfall event. The prediction quality of PLS models is rather low ($R^2 < 0.6$) but results can be useful to explore qualitatively the influence of rainfall on CSO characteristics.

Key words | canonical correlation, CSO, online monitoring, partial least squares regression, UV–Vis spectrometry

S. Sandoval
A. Torres
Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente,
Facultad de Ingeniería,
Pontificia Universidad Javeriana, Edificio J.G. Maldonado, S.J., Carrera 7 No. 40-62,
Bogotá,
Colombia

E. Pawlowsky-Reusing
Berliner Wasserbetriebe,
Netz- und Anlagenbau, Neue Jüdenstrasse 1, 10864 Berlin,
Germany

M. Riechel
N. Caradot (corresponding author)
Kompetenzzentrum Wasser Berlin gGmbH,
Cicerostraße 24, 10709 Berlin,
Germany
E-mail: *nicolas.caradot@kompetenz-wasser.de*

INTRODUCTION

The understanding of combined sewer overflow (CSO) pollutant dynamics requires detailed analyses of the variability and complexity of processes and interactions involved. The mentioned variability in wastewater and stormwater has already been addressed by multiple authors (e.g. Barraud *et al.* 2002; Veldkamp *et al.* 2002; Grüning & Orth 2002; Gruber *et al.* 2005; Hochedlinger *et al.* 2006; Bertrand-Krajewski *et al.* 2007; Aumond & Joannis 2008; Lacour *et al.* 2009; Schilperoort 2011; Métadier & Bertrand-Krajewski 2012). To better understand the processes and interactions related to this variability, different statistical methods can be applied (e.g. Grum & Aalderink 1999; Goodwin *et al.* 2003; Olyphant *et al.* 2003; Neuman *et al.* 2004; Even *et al.* 2007).

In the city of Berlin, a continuous integrated monitoring, using state-of-the-art online sensors, was started in 2010. It combined (i) continuous measurements of water quality and flow rates of CSO at one main CSO outlet and (ii) continuous measurements of water quality parameters at four sites within the urban stretch of the receiving river. Locally, the collection of data was also aimed at (i) characterizing CSO emissions, (ii) assessing the local dynamics and intensity of CSO impacts on the river and (iii) calibrating sewer and river water quality models as part of a planning tool for future CSO management in Berlin (Riechel *et al.* 2012). By means of characteristics of 22 online monitored rainfall and CSO events (e.g. total masses/volumes, mean and maximum concentrations and flow), multivariate statistical methods were implemented in order to (i) establish associations between rainfall and CSO characteristics (canonical correlation) and (ii) develop rainfall–CSO characteristics models (partial least squares (PLS) regression). The selected statistical methods could make it possible to: (i) identify which rainfall variable seems to be the most important in terms of prediction of each CSO characteristic; and (ii) verify the predictability of the models. Furthermore, uncertainties in the rainfall and CSO characteristics were assessed based on the variability of the PLS regression using the Monte Carlo method.

doi: 10.2166/wst.2013.524

MATERIALS AND METHODS

Study site

The CSO monitoring station was installed in a major overflow sewer about 230 m downstream of the main overflow structures and 500 m upstream of the outlet to the river. The station is located in a densely populated residential area with very few industrial activities. The catchment has ∼126,000 inhabitants, with a total impervious area of about 800 ha. The total area of the sub-catchment upstream of the outlet of the CSO monitoring is 100 ha and the impervious area of this sub-catchment is 74 ha. The sewer system is fully combined. Flow was measured directly in the overflow sewer, based on water level (air-ultrasound, Nivus company) and flow velocity (two ultrasound sensors, Nivus company). Water quality was measured on a bypass fed by a peristaltic pump, using (i) a UV–visible spectrometer (spectro::lyser, s::can company, Vienna) for the measurement of absorption water spectra which enable calculating equivalents of chemical oxygen demand (CODeq), dissolved COD (CODdeq) and total suspended solids (TSSeq) and (ii) a conductivity sensor (condu::lyser, s::can company, Vienna) for measurements of electrical conductivity (EC) and temperature. Rainfall measurements have been considered from four rain gauges (Figure 1) located around the CSO monitoring. Details about monitoring implementation can be found in Caradot (2012).

Figure 1 | Situation of the CSO monitoring station (white triangle) and of the four considered rain gauges: Berlin 9 (bln9), Spandau 2 (spa2), Charlottenburg 1 (chb1) and Wilmersdorf (wil).

CSO and rainfall characteristics

CSO and rainfall characteristics (see Table 1) as well as their uncertainties have been calculated for 22 events during the

Table 1 | CSO and rainfall characteristics for 22 events during the monitoring period

Rainfall characteristics

Charac.	Name	Unit	Mean	Min.	Max.
dur	Duration	min	236	45	965
imax_bln9	Max. intensity	mm/h	26.5	7.2	94.8
imax_wil	Max. intensity	mm/h	27.9	1.2	115.2
imax_spa2	Max. intensity	mm/h	23.5	4.8	73.2
imax_chb1	Max. intensity	mm/h	22.4	3.6	103.2
h_chb1	Depth	mm	10.4	2.3	25.6
h_bln9	Depth	mm	12.5	2.6	31.8
h_wil	Depth	mm	12.5	1.8	29.1
h_spa2	Depth	mm	12.7	1.5	34.9
imoy_chb1	Average intensity	mm/h	4.4	0.6	13.5
imoy_bln9	Average intensity	mm/h	4.6	1.1	11.8
imoy_wil	Average intensity	mm/h	5.0	0.2	14.6
imoy_spa2	Average intensity	mm/h	4.5	0.9	13.8
DWdur	Dry weather duration	day	3.9	0.0	18.9

CSO characteristics

Charac.	Name	Unit	Mean	Min.	Max.
dur	Duration	min	55	9	163
max_Q	Max. flow	m^3/s	1.6	0.6	3.5
V	Volume	m^3	3,452	304	12,375
mean_Q	Mean flow	m^3/s	1.0	0.4	2.0
M_COD	COD load	kg	1,186	68	4,003
M_TSS	TSS load	kg	784	46	2,765
M_CODd	CODd load	kg	209	12	775
mean_TSS	Mean TSS	mg/l	248	22	580
mean_COD	Mean COD	mg/l	377	35	908
mean_CODd	Mean CODd	mg/l	69	9	158
mean_EC	Mean EC	μS/cm	262	25	549
C_vol _waste	Wastewater ratio (volume)	%	11	0	34
C_COD _waste	Wastewater ratio (COD load)	%	16	0	46

monitoring period. Rainfall characteristics have been determined at the four considered rain gauges presented in Figure 1. Total measurements uncertainty has been estimated accounting for errors from sensor, lab chain, local calibration and field conditions (representativeness of site, see Métadier & Bertrand-Krajewski (2011)).

Over the 22 rainfall events, CSO volume ranges between 304 and 12,375 m³ and CSO COD load varies between 68 and 4,003 kg. Total COD load is 29 t ± 20%. COD mean concentration is 377 mg/l but it shows extreme variations from 35 to 908 mg/l. The average proportion of wastewater in the CSO (C_vol_waste) is 11% and the average contribution of wastewater to total CSO load (C_vol_waste) is 16%. This means that about 84% of the COD carried in the CSO originates from two other major sources: wash-off by stormwater runoff and resuspension of sewer sediments. For a full description of the CSO characteristics and calculation details, see Caradot *et al.* (2013).

Canonical correlation analysis

Relations between rainfall variables and CSO characteristics could be explored by means of multivariate statistical methods. Canonical correlation is a way of evaluating the linear relationship between two multidimensional data sets (Leurgans *et al.* 1993; González *et al.* 2008). The proposed method develops a number of functions that maximizes the correlation among canonical variates, which are orthogonal linear combinations of the original CSO and rainfall variables.

A couple of vectors a_1 and b_1 are found by maximizing the correlation function $corr(a_1^T X, b_1^T Y)$ between the two sets of variables X (input rainfall characteristics) and Y (output CSO characteristics), allowing the calculation of a first pair of canonical variates $CVx_1 = a_1^T X$ and $CVy_1 = b_1^T Y$. From this point, successive pairs of canonical variates CVx_i and CVy_i are computed with additional vectors a_i and b_i for explaining the residual correlation between the two sets of variables X and Y, i.e. the correlation between X and Y not accounted for by the first pair of canonical variates (Hair *et al.* 2005). Each additional pair of canonical variates CVx_i and CVy_i is orthogonal and independent of all other variates and the maximum number of canonical variates equals the number of variables in the smallest dataset, independent or dependent (Leurgans *et al.* 1993; González *et al.* 2008).

The canonical loadings (ρ) measure the linear correlations between the original variables and the canonical variates. These expressions can be computed as: $\rho = corr(CVx_i, X[:k])$ for a given rainfall characteristic k and the canonical variate CVx_i; and $\rho = corr(CVy_i, Y[:,j])$ for a given CSO characteristic j and the canonical variate $CVyi$. Therefore, the canonical loadings may lead to the identification of which rainfall variables are correlated with CSO characteristics, by analyzing the correlation of each individual variable over statistically significant (p-value <0.05) canonical variates (Hair *et al.* 2005).

PLS regression

PLS is a regression method widely used in chemistry, especially in chromatography and spectrometry applications (Tenenhaus 1998) where the number of variables (i.e. wavelengths) that characterize the spectra is very large compared to the number of observations, and where the prediction variables are highly correlated (Aji *et al.* 2003). This multicollinearity problem (and also the singularity ones) of the matrix composed by the independent variables is solved with PLS, by generalizing and fusing the principal component analysis and the multiple regression methods (Abdi 2003; Lee *et al.* 2011).

The PLS regression can integrate multiple dependent and independent variables (see Abdi 2003; Bertrand 2005). However, whenever more than one dependent variable is included (multivariate output), the PLS model loses accuracy, as well as predictability (Martínez Galera *et al.* 1997). Thus, multiple PLS models could be formulated by considering in all cases rainfall characteristics as the multivariate input, testing with different CSO characteristics as an individual output, adjusting the model parameters by a cross-validation procedure (e.g. a PLS model can be formulated with the following architecture: rainfall mean intensity, duration of the rainfall and total rainfall volume as three inputs and the TSS mean concentration as the unique output).

The PLS regressions can also bring up the identification of which variable of the multivariate independent variables set (rainfall characteristics) is the most representative over the specific output variable to be modeled (a given CSO characteristic), by the weights of the adjusted parameters of the regression (Rada Ariza & Torres 2011). Therefore, the PLS regression results to be obtained are: (i) the identification of the most important rainfall variable (related to the prediction of a CSO characteristic); and (ii) the predictability of the model (measured by the determination coefficient R^2). However, these results are sensitive to uncertainties, not only of the rainfall variables, but also of the CSO water quality/quantity characteristic to be predicted. Thus, the representativeness of these uncertainties in the results for (i) and (ii) (for any of the CSO characteristics) is assessed by the implementation of Monte Carlo simulations. The method consists of the repetitive generation of a set of random numbers that belong to the uncertainty interval of the different rainfall and CSO values (where each repetition will be a simulation). The results for (i) and (ii) can be estimated several times, according to the number of simulations, which can lead to identification of the number of times in which a specific rainfall characteristic was identified as the most important variable (probability of most important), as well as the variability and trend of the R^2 (adjustment of the PLS model), in terms of explaining a given CSO characteristic.

RESULTS AND DISCUSSION

Canonical correlation analysis

From the canonical correlation analysis, only two canonical variates (CV5 and CV6) appeared to be statistically significant. For both canonical variates, the correlation between CVx and CVy is close to one (e.g. $corr(a5^{T}X, b5^{T}Y) \sim 1$) so in the following CV refers to CVx = CVy. Table 2(a) shows the canonical loadings (ρ) obtained between CV5 and CV6 and the rainfall characteristics presented in Table 1 (independent variables). On the other hand, Table 2(b) shows the canonical loadings (ρ) obtained between canonical variates CV5 and CV6 and the CSO characteristics presented in Table 1 (response variables). From these tables, ρ values higher than 0.4–0.5 between CV5 and CV6 and both rainfall and CSO characteristics indicate possible influences of rainfall variables on CSO. Two complementary patterns can be identified:

(i) From the analysis of CV5, the maximum rainfall intensity and the average rainfall intensity seemed to have an influence on CSO quantity (volume, maximum and average flow rates) and CSO pollutant loads (TSS, COD and dissolved COD) (Table 2(a) and (b), in grey). Since the main CSO load contributions are the stormwater runoff and the resuspension of sewer sediments, the maximum rainfall intensity may influence the erosion and transport of particles in the wash-off.

(ii) From the analysis of CV6, the rain duration and the rain depth seemed to have an influence on CSO water quantity (duration and volume) and CSO water quality (mean concentration, EC and proportion of wastewater). The combination of rain duration and depth seems to be a good indicator to describe the CSO dilution and thus the contribution of wastewater to the CSO load.

These correlations results have been obtained using the maximum rainfall intensity calculated with a time step of 30

Table 2 | Canonical loadings (ρ) between (a) explanatory rainfall variables and (b) response variables with significant canonical variates CV5 and CV6

(a)	Correlations (canonical loadings ρ) between original *X* variables and significant canonical variates (CV)		(b)	Correlations (canonical loadings ρ) between original *Y* variables and significant canonical variates (CV)	
	CV5	CV6		CV5	CV6
dur	0.24	0.58	dur	0.00	0.64
imax_bln9	−0.43	0.35	max_Q	−0.54	0.39
imax_wil	−0.16	0.18	V	−0.40	0.72
imax_spa2	−0.45	0.49	mean_Q	−0.64	0.09
imax_chb1	−0.31	−0.34			
h_chb1	0.11	0.37	M_COD	−0.63	0.12
h_bln9	0.07	0.84	M_TSS	−0.51	0.23
h_wil	0.11	0.72	M_CODd	−0.72	0.13
h_spa2	−0.07	0.89			
imoy_chb1	−0.45	−0.45	mean_TSS	−0.24	−0.55
imoy_bln9	−0.52	−0.17	mean_COD	−0.41	−0.57
imoy_wil	−0.07	−0.25	mean_CODd	−0.47	−0.62
imoy_spa2	−0.66	0.06	mean_EC	−0.23	−0.71
DWdur	−0.21	0.22	C_vol_waste	−0.29	−0.71
			C_COD_waste	0.11	−0.60