Method for assessing the impact of rainfall depth on the stormwater volume in a sanitary sewage network

Edward Kozłowski¹ · Dariusz Kowalski² · Beata Kowalska² · Dariusz Mazurkiewicz³

Received: 28 August 2021 / Revised: 5 October 2021 / Accepted: 28 October 2021 / Published online: 12 November 2021
© The Author(s) 2021

Abstract

Sanitary sewage network is relatively rarely considered as the cause of urban floods. Its hydraulic overload can result not only in flooding, but also sanitary contamination of subcatchments. Stormwater is the main reason for this overload. In contrast to the stormwater or combined sewer system, these waters infiltrate into the network in an uncontrolled way, through ventilation holes of covers or structural faults and lack of tightness of manholes. Part of stormwater infiltrates into the soil, where it leaks into pipelines. This greatly hinders assessing the quantity of stormwater influent into the sanitary sewer system. Standard methods of finding correlation between rainfall and the intensity of stormwater flow are ineffective. This is confirmed, i.a. by the studies performed in an existing network, presented in this paper. Only when residuals analysis was performed using the ARIMA and ARIMAX methods, the authors were able to develop a mathematical model enabling to assess the influence of rainfall depth on the stormwater effluent from the sewage network. Owing to the possibility of using the rainfall depth forecasts, the developed mathematical model enables to prepare the local water and sewerage companies for the occurrence of urban floods as well as hydraulic overload of wastewater treatment plants.

Keywords Urban flooding · Sanitary sewage system · Stormwater influence · Estimation method · Prediction model

1 Introduction

Assessment of sanitary sewer flows is essential in the management of a wastewater disposal system. However, their prediction in real-time is extremely difficult due to a number

of factors affecting the flow rate, including different throughput (obstruction) of pipes, diversified water consumption patterns (sewage outflow) [1] and variability of meteorological phenomena. Determining the volume of rainwater collected is difficult and often requires specialized laboratory tests [2]. The assessment of this water inflow volume to the sanitary sewage system, which does not have specialized collecting devices, is even more difficult. In the literature, this problem is described only in the context of overflow from a combined sewer or stormwater sewer system during intense rainfall [3]. In a combined sewer system, the excess sewage and wet weather flows is discharged through stormwater overflow structures directly to the receiving water bodies, such as rivers or lakes, without prior treatment, contributing to their contamination [4, 5]. The overflows from sanitary sewers, which occur during intense rainfalls in small towns with no stormwater sewer system, constitute an even greater threat to the environment. If the municipal sewer system becomes significantly overloaded during intensive rainfalls, untreated combined sewage and stormwater flow out to the surface through manholes. This phenomenon is aggravated when the treatment capacity of a wastewater treatment plant (WWTP) is exceeded [6, 7]. Variations in flow of over two

✉ Dariusz Mazurkiewicz
d.mazurkiewicz@pollub.pl

Edward Kozłowski
e.kozlovski@pollub.pl

Dariusz Kowalski
d.kowalski@pollub.pl

Beata Kowalska
b.kowalska@pollub.pl

¹ Faculty of Management, Department of Quantitative Methods in Management, Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland

² Faculty of Environmental Engineering, Department of Water Supply and Wastewater Disposal, Lublin University of Technology, Nadbystrzycka 40B, 20-618 Lublin, Poland

³ Faculty of Mechanical Engineering, Department of Production Engineering, Lublin University of Technology, Nadbystrzycka 36, 20-618 Lublin, Poland

orders of magnitude are not uncommon in combined sewer systems [8]. Efficient sewage network management depends not only on accurate information on the momentary sewer flows but also the future expected flows, which enables to predict and avert emergencies [9, 10]. In addition, the possibility of predicting the hydraulic load of WWTP during extreme rainfalls enables to devise the algorithms for optimization of the treatment process [11].

Prediction of flows in particular locations of a combined sewer system is usually performed by means of hydrological models (rainfall/runoff) as well as computer simulation models. Unfortunately, such models require numerous input parameters which are necessary to describe a system discharging the volume of combined sanitary sewage and stormwater which is variable in time. Hence, model calibration is a difficult and time-consuming process. StormWater Management Model (SWMM), devised by the United States Environmental Protection Agency (US EPA) constitutes a popular tool for simulating the flow of combined sanitary sewage and stormwater [12]. This is open source computer software, enabling integration with an external real-time control module. Connecting the computer data with real-time control system (RTC) requires installation of sensors and actuators in the network, enabling to monitor the flow of sewage periodically supplied with stormwater [13, 14]. These data are collected by sensors and sent to the control component, which calculates an optimized sewer system control strategy. The majority of RTC implementations are limited to local reactive or supervisory control. However, real-time control is more efficient, if it is integrated with the entire sewer system. Pleau et al. [15] employed a sewer networks global optimal control (GOC) scheme with a two-level architecture, in which the upper level was composed of a central station, whereas the lower level is composed of local stations. The real-time computer is dedicated to all RTC operations. In the optimizing algorithm, they defined a multi-objective (cost) function and a set of constraints, based on the following control objectives: minimizing overflows, minimizing set point variations and maximizing the use of WWTP capacity. The GOC approach was also proposed by Fu et al. [16] who employed a genetic algorithm for the optimization of the multi-objective function. Schütze et al. [17] proposed a real-time global optimal control system in which the control objectives involve minimization of overflows, the maximization of the use of the treatment plant capacity, the minimization of accumulated volumes and, the minimization of variations of the setpoints. Beeneken et al. [18] employed a global real-time control system which was implemented in the combined sewer network in Dresden, Germany. To allow for the realization of a global RTC, the existing process control system (PCS) was provided with an additional control computer connected to the SCADA system. Darsono and Labadie [19] proposed RTC based on a neural optimal

algorithm integrating a dynamic hydraulic model with an optimizing model, aimed at minimizing overflows and simultaneously maximizing through-flows to the WWTP for a storm event. On the basis of RTC, Vezzaro and Grum [20] employed the Dynamic Overflow Risk Assessment (DORA) approach in the Lynetten catchment (Denmark) to minimize combined sewer overflow (CSO) volumes, which aimed at reducing overflow risk in the system by minimizing a global cost function through an optimization routine. Garofalo et al. [5] presented a method of flow control in a municipal combined sewer network which involved installing sensors and a series of electronically movable gates controlled by a decentralized real-time system based on a gossip-based algorithm, integrated with the hydrodynamic simulation model, SWMM. This method was implemented in the municipal sewer system of Cosenza (Italy). As a result, it was possible to take advantage of the sewer network capacity and accumulate excess combined sewage and stormwater in the locations, where pipes were underloaded and there was no risk of combined sewer overflow (CSO). Chen et al. [9] devised a real-time future flow prediction algorithm based on autoregressive moving average (ARMA) models and multistep iterative prediction. Then, online control of chemical dosing was implemented using the predicted flow. An Event-driven Model Predictive Control (EMPC) methodology was presented in the work by Liu et al. [10] who proposed controlling the flows of sewage streams containing the dosed chemical. Li et al. [21] employed autoregressive with exogenous inputs (ARX) models to reduce the delays with rainfall data used as model inputs to control chemical dosing in sewer systems. Using a hydraulic model and Long–Short-Term Memory LSTM, Zhang et al. [22] presented a new Inter Catchment Wastewater Transfer method which can mitigate sewer overflows. Jean et al. [23] evaluates the effects of three rainfall data selection methods on the estimation of CSO. The methodology involved hydrological/hydraulic modeling of an urban catchment in the Province of Québec (Canada). Continuous simulation provided the most accurate volume estimations and showed high sensitivity to the number of simulated years.

Another approach to the issue of flow prediction is the application of “black box” models, which include artificial neural network (ANN) models. In these models, the internal functions are not described; instead the relations between measurable input and output data are relevant. One of the first studies which aimed at the application of data-driven methods were performed by Carstensen et al. [24]. The authors predicted the hydraulic load of a WWTP an hour in advance using a simple regression model, adaptive “gray box” model (comprising a simple deterministic regression model and a stochastic model component) as well as a complex hydrologic model, which was calibrated on the basis of extensive measurement campaigns in a sewer system.

Darsono and Labadie [19] proposed the neural–optimal control algorithm in a simulated real-time control experiment for the King County combined sewer system, Seattle, Washington, USA. Optimal control model was utilized to provide the training data set for a recurrent artificial neural network (ANN) under a wide range of sewer inflow conditions. El-Din and Smith [25] presented the possibilities for ANN application to make short-term predictions of wastewater inflow rate that entered the Gold Bar Wastewater Treatment Plant (GBWWTP), the largest plant in the Edmonton area (Alberta, Canada). The neural model used rainfall data, observed in the sewage system discharging to the plant, as inputs. Wei et al. [26] employed a multi-layer perceptron neural network algorithm to predict flow at the treatment plant at Des Moines Wastewater Reclamation Facility (WRF) in Iowa, USA. In addition to the current rainfall data, they also used radar reflectivity data to achieve an improved flow prediction for different short time horizons. Zhang et al. [22] used different recurrent neural networks for flow prediction in a sewer system in Drammen, Norway. First, a hydraulic model was employed to identify relatively dry pipes (control target), then they used RNNs (Recurrent Neural Network) to realize flow prediction for the target pipe. Three RNN architectures, namely, Elman, NARX (nonlinear autoregressive network with exogenous inputs) and a novel architecture of neural networks, LSTM (Long–Short-Term Memory), were compared. Zhang et al. [27] applied different neural networks for predicting the water level of a combined sewer overflow using the data collected from an Internet of Things monitoring CSO structure. Karimi et al. [3] used and then compared three data-driven methods to predict flow: (1) artificial neural network (ANN); (2) long–short term memory (LSTM); and (3) least absolute shrinkage and selection operator (LASSO). The analyses were performed using the data from a sewer system monitoring in Springfield, USA. Although all three data-driven methodologies ensure acceptable prediction performance, they observed that LSTM surpasses ANN, while the LASSO method both ensures good prediction performance and aids in identifying the key parameters which affect the flow in any given location of the network. During flow prediction, the authors also accounted for the amount of groundwater as additional input data in all three methods. To better predict the flows corresponding to rare events, such as single rainfall event per 50 or 100 years, they employed the resampling approach (SmoteR) to modify the training data set [3].

The aim of the paper was to present the prediction model of sewage inflow from sanitary sewer to a wastewater treatment plant, accounting for the stormwater and infiltration inflow. The novelty of the article consists in the presentation of the process involving search for an appropriate description method, which could be employed for a sanitary network susceptible to the inflow of stormwater and

simultaneously devoid of specialized devices for their interception. The first stage of the search included studying the seasonality of sewage inflow rate to a WWTP, and then the likelihood method of Ljung–Box test and linear regression were used to evaluate the necessity of accounting for the time lag between rainfall and inflow to the WWTP and the influence of stormwater accumulation over several days of rainfall on the inflow volume. The unsatisfactory fit of the expected value to the measured one led the authors to the application of ARIMA and ARIMAX models which enabled to achieve the required fit. The goodness of fit comparison for the considered models was based on the seasonality and integration tests as well as Akaike's Information Criterion (AIC).

2 Description of the object

The considered sanitary sewer system is located in a mountainous area, in southern Poland. It serves approximately 40 thousand people, including 30 thousand in the city and 10 thousand in the adjacent villages. Both in the urban and rural areas, there are no large industrial plants discharging wastewater to the sewer network. The constructed system mainly discharges municipal sewage as well as household sewage from civic buildings, services and small industrial plants. No stormwater sewer system was built in the considered area.

Total length of pipes in the sewer network approximates 1200 km. It was mainly built of PVC (55%) and stoneware pipes (38%). The age of pipes is diversified, ranging from 10 to 70 years. The pipes which are newer than 20 years constitute about 80% of the entire network, whereas older ones—approximately 20%. In line with the design intent, the analyzed network was not equipped with the devices stormwater interception. There are no storm inlets or stormwater overflow structures.

The wastewater collected by the sewer system are discharged to a collective wastewater treatment plant with the capacity of 40 000 m³/d, located within the city zone (Fig. 1). A flow meter measuring the wastewater flow rate is located at the WWTP inlet. Gravity wastewater flow is dominant in the considered sewer network; however, due to the mountainous location, it was necessary to use 98 network pumping stations.

The mean daily volume of sanitary wastewater influent to the sewer system, estimated on the basis of the volume of sold water, reached approximately 7 800 m³/day in 2019. Therefore, the considered WWTP has a substantial reserve in capacity. The maximum time of sewer flow from pluviometers to the treatment plant, estimated with SWMM 5.0 software, is 4 days. Other realised simulation studies conducted using this software, based on the daily water consumption indicated that the sanitary sewer

Fig. 1 Scheme of the sewer network with the marked location of the wastewater treatment plant and pluviometers

network also has a large capacity reserve, taking into account the inflow of sanitary sewage. Unfortunately, due to a lack of stormwater sewer system in the considered area, combined stormwater and infiltration flow reach the sanitary sewer system in an uncontrolled way. The adopted project assumptions assuming 100% capacity reserve of the constructed sanitary sewer system turned out to be insufficient. In 2019 alone, 26 wastewater overflows were observed due to the hydraulic overload. In all cases, the overflow occurred in more than 1 location, involving from 4 to 18 points at a time. The areas in proximity of pumping stations were flooded most often; however, floodings also occurred from the manholes in direct adjacency of residential buildings. The emergency overflow structure at the wastewater treatment plant was triggered four times, discharging untreated wastewater to the river. The simulation studies carried out for the needs of a local company operating the sewer system indicated that when 17,000 m³/day is exceeded there is a risk that the wastewater from sewer pipes will flow onto the surface.

Therefore, the local water company took the steps connected with devising a system warning against the hydraulic overload of both the sewer system and the wastewater treatment plant. This system is based mainly on the installation of 2 pluviometers (Fig. 1) as well as creation of a method for predicting the inflow of wastewater to the WWTP, based on the historic data of rainfall depth and the measured wastewater inflow. Unfortunately, the prediction model employed by the water company was ineffective. Therefore, it was necessary to devise a new model.

3 Methodology

The wastewater prediction model was based on the read-outs of two automatic RG13 ENVAG trough-type pluviometers as well as an ABB Venturi sewage flow meter. The location of measurement devices is presented in Fig. 1. Both the flow meter and pluviometers sent the measurement data to the sewage system dispatcher via the SCADA system. The measurement data made available to the authors of this paper comprised the daily rainfall depth (mm/day) and the daily volume of wastewater (m³/day) flowing into the plant. The measurement data corresponded to the period from 1st January 2019 to 30th April 2020. The model was created in the following steps:

- conducting a seasonality study, which aimed at verifying whether increased sewage inflow occurs in the considered sewage network on particular days—assessment of the influence of service and industrial facilities on the sewage inflow volume,
- verifying the necessity of accounting for time lag between rainfall and inflow to the WWTP (0–3 days), in the created prediction model,
- verifying the influence of rainfall accumulation (1–4 days) on the volume of sewage inflow to a WWTP,
- creating an ARIMA prediction model and subsequently ARIMAX model accounting for the influence of rainfall depth as well as additional, unidentified factors on the volume of sewage inflow to a WWTP.

3.1 Seasonality study

Let $\{y_i\}_{1 \leq i \leq n}$ denote a sequence representing the volume of sewage flowing into the plant, $y_i \in R_+$ for $1 \leq i \leq n$ and R_+ denotes the set of non-negative real numbers. The sequence $\{y_i\}_{1 \leq i \leq n}$ we can present a set of sequences $\{z_k\}_{1 \leq k \leq 7}$ corresponding to consecutive groups—days of the week (1—Sunday, 2—Monday, ..., 7—Saturday), where the subsequence $z_k = \{y_{k+7j}\}_{1 \leq j \leq n_k}$ contains the data on the sewage inflow volume for k th group, n_k number of elements (cardinality) in k th group and $n = n_1 + n_2 + \dots + n_7$. Each of the sequences corresponds to the phase occurring in time series [1]. If the sequences meet the normality assumption, analysis of variance (ANOVA) method can be employed for determining the significant difference in groups; conversely, if the normality assumption is not met, the Kruskal–Wallis test or Friedman's test are used see, e.g., [28]. In the paper, Kruskal–Wallis was applied for verifying seasonality was performed using a Kruskal–Wallis test. At the significance level of $\alpha \in (0,1)$ a null hypothesis was created:

H_0 : distributions of sewage inflow volume for each day of the week are equal/identical (the day of the week does not affect the volume of sewage).

In addition, an alternative hypothesis was formulated as follows:

H_1 : there is at least 1 day for which the sewage volume distribution is significantly different from the rest.

Ranking is performed for the entire $\{y_i\}_{1 \leq i \leq n}$ sample. Let R_{ij} denote the rank of the j th element from the i th group within the sample.

The test statistic is expressed with the formula:

$$T = \frac{12}{n(n+1)} \sum_{i=1}^7 \left(\bar{R}_i - \frac{n+1}{2} \right)^2 n_i, \quad (1)$$

where

$$\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}. \quad (2)$$

The statistic T has χ^2 distribution with 6 degrees of freedom. It is a measure of deviation of mean ranks in groups from the mean value of all ranks, equal to $(n+1)/2$. The value $\chi^2_{1-\alpha,6}$ denotes a quantile of $1-\alpha$ order for χ^2 distribution with 6 degrees of freedom. If $T < \chi^2_{1-\alpha,6}$ then at the significance level α there is no grounds to reject the null hypothesis H_0 , otherwise the hypothesis H_0 should be rejected in favor of the alternative hypothesis H_1 . The investigations were conducted using R-Languages Programming [29].

3.2 Maximum likelihood estimation

Let $\{y_i\}_{1 \leq i \leq n}$ denote the sequence of dependent variable as well as $y_i \in \mathbb{R}$ for $1 \leq i \leq n$, whereas $\{x_i\}_{1 \leq i \leq n}$ denotes the sequence of predictor values and $x_i \in \mathbb{R}^k$. The dependence between the response variance and predictors can be described using the equation:

$$y_i = f(\theta, x_i) + \varepsilon_i, \quad (3)$$

where $\theta \in \mathbb{R}^{k+1}$ denotes the unknown model parameters (it was assumed that an intercept in additionally exists in the model), $\{\varepsilon_i\}_{1 \leq i \leq n}$ is a sequence of independent random variables with normal distribution $N(0, \sigma^2)$. In the paper maximum likelihood estimation (MLE) method [30, 31] was used to determine the unknown parameters θ and standard deviation values σ . For this purpose the likelihood function is created as follows:

$$L(\theta, \sigma) = \prod_{i=1}^n \gamma(y_i, f(\theta, x_i), \sigma), \quad (4)$$

where

$$\gamma(s, m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right), \quad (5)$$

is a density function of normal distribution $N(0, \sigma^2)$. By maximizing the logarithm of the likelihood function and solving the task:

$$\max_{\theta, \sigma} \ln L(\theta, \sigma), \quad (6)$$

we determine the estimators of unknown parameters θ and standard deviation σ . The following indices were used for comparison of model quality:

- Coefficient of determination $R^2 = 1 - \frac{\sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$.
- Akaike's Information Criterion (AIC) $AIC = -2\ln L(\theta, \sigma) + 2(k+1)$.
- Root mean square error (RMSE) $RMSE = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$.
- Mean absolute error (MAE) $MAE = \frac{1}{n} \sum_{i=1}^n |\varepsilon_i|$.
- Mean absolute percentage error (MAPE) $MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\varepsilon_i}{y_i} \right|$.

For linear regression model the value $R^2 \in (0,1)$ and when R^2 is closer to 1, then the model (3) is better fitted to the observed (response variable and predictors) data. Whereas the lower the value of AIC, RMSE, MAE, MAPE indices designate the better fitting of model to the data.

3.3 Ljung–Box test

Verification of independence of the residual sequence $\{\varepsilon_t\}_{1 \leq t \leq n}$ in models (3) was performed using the Ljung–Box test [31, 32]. At significance level $0 < \alpha < 1$ we create a null hypothesis as follows:

H_0 : the elements of residual sequence $\{\varepsilon_t\}_{1 \leq t \leq n}$ are independent (no correlation between the sequence elements); and an alternative hypothesis: H_1 : the elements of residual sequence $\{\varepsilon_t\}_{1 \leq t \leq n}$ are dependent (correlation occurs between the sequence elements for shifts $1 \leq \tau \leq s$, the correlation coefficient is significantly different from 0).

The test statistic:

$$Q = n(n+2) \sum_{\tau=1}^s \frac{r_\tau^2}{n-\tau}, \quad (7)$$

has χ^2 distribution with s degrees of freedom, whereas r_τ denotes the value of correlation coefficient between $\{\varepsilon_t\}_{1 \leq t \leq n}$ series elements with shift τ . If $Q < \chi_{1-\alpha, s}^2$ ($\chi_{1-\alpha, s}^2$ —quantile of order $1 - \alpha$ for χ^2 distribution with s degrees of freedom), then at the significance level α there is no basis to reject the null hypothesis H_0 , otherwise the null hypothesis H_0 should be rejected in favor of the alternative hypothesis H_1 .

3.4 Influence of time lag between rainfall and sewage inflow

The influence of rainfall levels and time lag on the sewage inflow was assessed by means of linear regression:

$$y_t = \beta_0 + \beta_1 g_{t-\tau} + \beta_2 k_{t-\tau} + \varepsilon_t, \quad (8)$$

where $\{g_t\}_{1 \leq t \leq n}$ and $\{k_t\}_{1 \leq t \leq n}$ denote the sequences representing the rainfall levels in the measurement locations Pluviometer 1 and Pluviometer 2, t —time, τ —time lag (0–3 days), $\{\varepsilon_t\}_{1 \leq t \leq n}$ is a sequence of residuals. Identification of unknown parameters $\beta_0, \beta_1, \beta_2$ in model (8) was performed with the MLE. In addition, the correlation of residual sequence $\{\varepsilon_t\}_{1 \leq t \leq n}$ was verified by means of the Ljung–Box test.

3.5 Influence of rainfall accumulation on sewage inflow

To evaluate the influence of rainfall accumulation over consecutive days on the sewage inflow to the WWTP, the following model was employed:

$$y_t = \beta_0 + \beta_1 \sum_{j=0}^{\tau} g_{t-j} + \beta_2 \sum_{j=0}^{\tau} k_{t-j} + \varepsilon_t \quad (9)$$

where $\sum_{j=0}^{\tau} g_{t-j}$ and $\sum_{j=0}^{\tau} k_{t-j}$ denote the accumulated rainfall depth measured by means of pluviometers 1 and 2, over the period from the moment $t - \tau$ to the moment t . The rainfall depth accumulation over the period of 1–4 days was assumed in the calculations according the estimated (SWMM 5.0) time of sewage flow. Identification of unknown parameters $\beta_0, \beta_1, \beta_2$ in model (9) was performed with the MLE, whereas the correlation of residual sequence $\{\varepsilon_t\}_{1 \leq t \leq n}$ was carried out with the Ljung–Box test.

3.6 ARIMA and ARIMAX prediction models

Regression models (8) and (9) enable the assessment of time lag and rainfall depth accumulation on the sewage inflow to a WWTP. Nevertheless, the existence of residual sequence correlation in these models indicates the existence of additional unknown factors affecting the volume of sewage inflow. The presence of such factors is connected with the infiltration process, sometimes with the flow of sewage from the sewage system to a WWTP and uncontrolled inflow of stormwater to the sewage system—lack of manhole tightness, manhole cover vents, illegally removed manhole covers, illegal roof drains. To account for the influence of above-mentioned factors in the sewage inflow prediction model, the authors decided to employ an Autoregressive Integrated Moving Average (ARIMA) model [32–35].

Identification of sewage inflow to the WWTP was carried out with the ARIMA(p, r, q) model in the following form:

$$\Delta^r \hat{y}_t = \alpha_1 \Delta^r \hat{y}_{t-1} + \dots + \alpha_p \Delta^r \hat{y}_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}, \quad (10)$$

where $\hat{y}_t = y_t - \bar{y}$, difference operator Δ is expressed with the formula $\Delta \hat{y}_t = \hat{y}_t - \hat{y}_{t-1}$ and difference of order $r > 1$ is determined as $\Delta^r \hat{y}_t = \Delta^{r-1} \hat{y}_t - \Delta^{r-1} \hat{y}_{t-1}$, whereas $\{\varepsilon_t\}_{1 \leq t \leq n}$ is a sequence of independent random variables with normal distribution $N(0, \sigma^2)$. In ARIMA(p, r, q) model $p, r, q \in N$ (N —set of non-negative integers) and p denotes the autoregression order, q —the moving average order, r —the degree of differencing. The estimators of parameters $\alpha_1, \dots, \alpha_p, \theta_1, \dots, \theta_q$ of model (10) and the estimator of standard deviation σ were determined using the MLE.

Equation (10) can also be presented in the form:

$$\Phi_p(B)(1-B)^r(y_t - \bar{y}) = \Theta_q(B)\varepsilon_t, \quad (11)$$

where

$$\Phi_p(B) = (1 - \alpha_1 B + \dots + \alpha_p B^p), \quad (12)$$

$$\Theta_q(B) = (1 - \theta_1 B - \dots - \theta_q B^q), \quad (13)$$

and the shift operator B is determined as $B^k y_t = y_{t-k}$ for $t > k$. From a set of possible ARIMA models we select such ARIMA(p, q, r) for which AIC is the lowest.

The model based on formula (10) does not take into account the amount of rainfall. Assuming that this value has a significant impact on the amount of sewage flowing to the WWTP, the authors decided to use the ARIMAX class model, in which they included rainfall accumulation over consecutive days as regressors. The volume of sewage influent to the WWTP was also determined using the ARIMAX(p, r, q) model in the following form [31–33]:

$$y_t = \beta_0 + \beta_1 \sum_{j=0}^{\tau} g_{t-j} + \beta_2 \sum_{j=0}^{\tau} k_{t-j} + \varepsilon_t, \quad (14)$$

$$\Delta^r \varepsilon_t = \alpha_1 \Delta^r \varepsilon_{t-1} + \dots + \alpha_p \Delta^r \varepsilon_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}, \quad (15)$$

where the designations used are the same as in the previous formulae. The ARIMAX (p, r, q) is an expansion of ARIMA (p, r, q) models, accounting for the influence of predictors on the dependent variable. The system of Eqs. (14)–(15) can also be presented in the form:

$$\Phi_p(B)(1-B)^r \left(y_t - \beta_0 - \beta_1 \sum_{j=0}^{\tau} g_{t-j} - \beta_2 \sum_{j=0}^{\tau} k_{t-j} \right) = \Theta_q(B) \varepsilon_t. \quad (16)$$

Similarly, from a set of possible ARIMAX models we choose the model for which AIC is the lowest.

4 Results and discussion

Figure 2 presents the graph with the daily rainfall depth indicated by the pluviometers 1 and 2 as well as the daily volume of sewage influent to the WWTP. It shows the lack of 30-day periodicity.

The graphs presented in Fig. 2 are insufficient to draw conclusions on the influence of rainfall depth on the volume of sewage inflow to the WWTP. Determining this influence requires further statistical analyses.

4.1 Seasonality study

First, the normal distribution hypothesis in each group was verified with the Shapiro–Wilk normality test. The p value was determined for each tested group; this value expresses the likelihood of obtaining the test results which are at least as extreme as the results actually observed, provided that

Fig. 2 Daily rainfall depth and volume of sewage influent to the WWTP over the study period

Table 1 *P* values for each of the analyzed groups

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
3.528×10^{-7}	8.698×10^{-9}	4.706×10^{-9}	5.046×10^{-12}	1.075×10^{-11}	3.866×10^{-10}	1.924×10^{-6}

Fig. 3 Box-plot graph (min, max, outliers, quantiles) for each day of the week **Fig. 4** Graph showing the volume of sewage for each day of the week throughout the entire study period

the null hypothesis is correct. The *p* value for each group is presented in Table 1.

Therefore, at the significance level of 0.01 for each group, the normal distribution hypothesis should be rejected in favor of the alternative hypothesis. Therefore, the seasonality of sewage inflow was studied by means of the Kruskal–Wallis test.

The value of test statistic *T* (1) estimated based on the Kruskal–Wallis test amounts to 8.982, whereas *p* value = 0.1746 (*p* value > 0.01). Hence, there were no grounds to reject the null hypothesis and it can be assumed that the difference between the sewage volume distribution depending on the day of the week is insignificant. Therefore, for the analyzed sewer system, the influence of the day of the week is irrelevant. This conclusion is confirmed by the

analysis of the figures below showing the sewage volume on each day of the week (Figs. 3, 4).

Taking into account the presented analysis results, it can be stated that no seasonality is observed in the analyzed sewer system, in relation to the days of the week. No significant trends were observed either. This confirms the results lack of large industrial plants as well as negligible influence of the functioning services and small industrial plants. Due to the fact that seasonality does not affect the sewage volume; hence, this factor can be omitted in regression analyses.

4.2 Influence of time lag between rainfall and sewage inflow

Table 2 presents the values of structural parameter estimators for models (8) accounting for linear regression of sewage volume on rainfall level with the lag $\tau = 0, 1, 2, 3$ days. In addition, the standard deviation of these parameters, the value of *T* statistic and *p* value for the significance test of structural parameters, the value of *Q* statistic (7), and *p* value for the Ljung–Box test were given. This table also contains the evaluation of goodness of fit between the models and values of empirical data, including: coefficient of determination R^2 , AIC, values of RMSE, MAE, and MAPE.

The goodness of fit obtained on the basis of the linear regression model is unsatisfactory, regardless of the assumed lag of sewage inflow to the WWTP. In each case, the coefficient of determination R^2 is lower than 0.5. The lowest AIC value, indicating goodness of fit, was obtained for lag $\tau = 1$. The lowest values of remaining indices (RMSE, MAE and MAPE) were obtained with no lag ($\tau = 0$). The insufficient goodness of fit between the results of model calculations and measured values is shown in

Table 2 Summary of identification results of linear model (8), accounting for the lag between rainfall event and sewage inflow

Lag, days	$\tau=0$	$\tau=1$	$\tau=2$	$\tau=3$
β_0	20,566.886	20,613.288	20,781.293	20,919.164
S_{β_0}	263.245	267.954	278.965	287
T_{β_0}	78.128	76.928	74.494	72.889
$p.val_{\beta_0}$	3.43×10^{-276}	7.99×10^{-273}	2.89×10^{-266}	9.64×10^{-262}
β_1	392.627	380.712	321.42	296.72
S_{β_1}	61.372	62.398	65.961	67.783
T_{β_1}	6.398	6.101	4.873	4.377
$p.val_{\beta_1}$	3.736×10^{-10}	2.162×10^{-9}	1.496×10^{-6}	1.475×10^{-5}
β_2	132.687	116.106	99.48	56.061
S_{β_2}	55.847	56.78	59.356	60.994
T_{β_2}	2.376	2.045	1.676	0.919
$p.val_{\beta_2}$	0.0179	0.0414	0.0944	0.3585
R^2	0.251	0.225	0.159	0.113
AIC	9744.71	9740.732	9759.627	9765.767
RMSE	5438.992	5529.807	5756.964	5915.815
MAE	3606.754	3781.055	3871.534	3939.33
MAPE	0.157	0.166	0.169	0.172
χ^2	1124.783	991.223	877.635	865.91
$p.val_{\chi^2}$	0	0	0	0

Fig. 5 Comparison of measured and predicted volume of sewage estimated by means of model (8) for lag $\tau=1$. Calculation step—1 day

Fig. 5. It presents the comparison of calculation results for lag $\tau=1$ day—for which the AIC value was the lowest.

Taking into account the presented results, it should be stated that the goodness of fit of model (8) to the empirical data—for the considered lag between a rainfall event and sewage inflow amounting to $\tau=0,1,2,3$ days—is insufficient. However, it was noted that the rainfall levels significantly affect the amount of sewage influent to the WWTP ($p.val_{\beta_1} < 0.01$ for $\tau=0,1,2,3$ and $p.val_{\beta_2} < 0.05$ only for $\tau=0,1$). In addition, it was observed that there is a significant correlation in the residual sequence $\{\epsilon_t\}_{\tau+1 \leq t \leq n}$, which

suggests that there are factors other than rainfall which affect the volume of sewage flowing into the WWTP.

4.3 Influence of rainfall accumulation on the sewage inflow

Table 3 presents the values of estimators of structural parameters for models (8), accounting for the rainfall depth accumulation from consecutive $\tau=1,2,3,4$ days. In addition, the standard deviation of these parameters, values of the T

Table 3 Summary of identification results of linear model (8) accounting for the rainfall accumulation from consecutive days $\tau = 1, 2, 3, 4$

Rainfall depth accumulation, days	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$
β_0	19,895.624	19,430.804	19,063.048	18,785.506
S_{β_0}	247.775	240.719	239.18	245.211
T_{β_0}	80.297	80.72	79.702	76.61
$p.val_{\beta_0}$	3.79×10^{-281}	8.62×10^{-282}	6.00×10^{-279}	6.12×10^{-271}
β_1	316.397	265.551	230.086	213.631
S_{β_1}	38.319	30.072	25.592	23.388
T_{β_1}	8.257	8.831	8.991	9.134
$p.val_{\beta_1}$	1.448×10^{-15}	1.950×10^{-17}	5.686×10^{-18}	1.858×10^{-18}
β_2	95.581	78.704	69.451	50.941
S_{β_2}	35.396	28.149	24.418	22.562
T_{β_2}	2.7	2.796	2.844	2.258
$p.val_{\beta_2}$	0.00717	0.0054	0.0047	0.0244
R^2	0.383	0.45	0.487	0.491
AIC	9630.556	9554.179	9501.515	9478.336
RMSE	4936.068	4656.063	4499.997	4483.608
MAE	3299.782	3232.249	3165.604	3115.744
MAPE	0.145	0.146	0.145	0.144
χ^2	1234.941	1258.127	1238.733	1123.676
$p.val_{\chi^2}$	0	0	0	0

Fig. 6 Comparison of the measured and predicted sewage volume estimated with the model (8) for rainfall depth accumulation from $\tau = 4$ days. Calculation step—1 day

statistic and p value for significance test of structural parameters, value of the Q statistic and p value for the Ljung–Box test. This table also contains the goodness of fit of the model to empirical data.

Figure 6 additionally presents the comparison between the measured and calculated volume of sewage influent to the WWTP. The comparison involves a model accounting for 4-day rainfall depth accumulation, for which AIC and the values of other indices (RMSE, MAE, and MAPE) were the lowest. Accounting for the accumulated rainfall depth in the linear regression model (8) slightly improved

the goodness of fit between the calculated and measured values. Coefficient of determination increased up to 0.491. In turn, AIC and the values of other indices (RMSE, MAE, and MAPE) decreased slightly. Similarly as in the previous case, the influence of rainfall depth accumulation on the volume of sewage influent to the WWTP was deemed significant. A significant correlation was also observed in the residual sequence $\{\varepsilon_t\}_{\tau+1 \leq t \leq n}$, suggesting the existence of factors other than rainfall, which affect the volume of sewage flowing into the WWTP. However, the goodness of

Table 4 Parameters of the ARIMA (1, 0, 1) model

	Estimator	Std. error	T	p value
α_1	0.8005	0.0324	24.7113	0
θ_1	0.1042	0.0541	1.9272	0.0545
$\bar{y}$	21,773.4222	860.0352	25.3169	0

Table 5 Parameters of the ARIMAX (4, 1, 2) model

	Estimator	Std. error	T	p value
α_1	-0.4449	0.0920	-4.8368	1.79×10^{-6}
α_2	0.2218	0.0903	2.4568	0.0144
α_3	-0.2659	0.0581	-4.5724	6.16×10^{-6}
α_4	-0.2507	0.0517	-4.8531	1.65×10^{-6}
θ_1	0.1969	0.0894	2.2026	0.0281
θ_2	-0.5750	0.0861	-6.6759	6.89×10^{-11}
β_1	200.4697	20.7442	9.6639	0
β_2	33.5989	20.6309	1.6286	0.104

Table 6 Comparison of indices in ARIMA (1, 0, 1) and ARIMAX (4, 1, 2) models

	AIC	RMSE	MAE	MAPE
ARIMA (1, 0, 1)	9307.54	3464.612	2073.631	0.089
ARIMAX (4, 1, 2)	9025.26	2840.392	1826.517	0.080

fit was still insufficient for predicting the sewage volume, required for controlling the sewer system operation.

4.4 ARIMA and ARIMAX prediction models

Out of the ARIMA models (10) the best goodness of fit to the empirical data was obtained for the ARIMA (1, 0, 1). Table 4 presents the values of estimators, standard deviations of these parameters as well as the values of the T statistic and p value enabling to assess the significance of parameters in this model.

Values of AIC for ARIMA (1, 0, 1) amounts to 9307.54. It is much lower than that obtained using the models (8) and (8). Additional indices describing goodness of fit are presented in Table 6.

In the case of ARIMAX models (16), the best fit to the empirical data was obtained for the ARIMAX (4, 1, 2) model taking into account the rainfall accumulation over 4 days ($\tau=4$). Table 5 presents the values of estimators, standard deviations of these parameters as well as the values of the T statistic and p value enabling to assess the significance of parameters in this model.

The AIC value for the ARIMAX (4, 1, 2) model is equal to 9025.26. This value is much lower than the results obtained using regression models (8) and (8), as well as the ARIMA (1, 0, 1) model. Much better fit was also confirmed by the values of other indices, as presented in Table 6.

Fig. 7 Comparison of the sewage volume measured and predicted by means of ARIMA (1, 0, 1) and ARIMAX (4, 1, 2) models (16). Calculation step—1 day

Figure 7 presents the graphs of daily sewage volume predicted by means of ARIMA (1, 0, 1) and ARIMAX (4, 1, 2) models as well as the volumes measured at the WWTP inlet.

The ARIMA and ARIMAX models achieved much better fit to the observed wastewater values than the regression models presented earlier. The best fit was obtained while using the ARIMAX (4, 1, 2) model. This model can be employed for estimating the volume of sewage for the purpose of controlling the WWTP operation and prepare the operational teams for the expected sewage overflows from the sewer system.

5 Conclusions

Taking into account the conducted studies, it should be noted that the issue of uncontrolled inflow of stormwater to the sanitary sewer system is significant and may cause a hydraulic overload of both the network and WWTP. This inflow can be difficult to estimate, even with meteorological stations measuring the rainfall depth as well as flow meters installed within the sewer network.

Nevertheless, correct management of a sanitary sewer system and a WWTP requires the knowledge of the influent sewage volume, which is especially important in the situations, where their capacity might be exceeded. The predicted volume of influent sewage constitutes a basic element of such management, which necessitates devising and implementing an appropriate mathematical model, calibrated on the basis of historical data.

The paper showed that the mathematical models based on linear regression are inadequate for predicting the sewage volume. This mainly results from the uncontrolled nature of stormwater inflow to the sanitary sewer system. Classic methods of time series decomposition, such as STL—seasonal and trend decomposition using Loess or exponential smoothing do not take into account the external factor, such as the amount of rainfall, which causes an increase in the amount of sewage flowing into the sewage system. Nevertheless, the amount of wastewater is correlated with the amount of rainfall. That is why the ARIMA models were employed in the article enable to account for these pathways, even without the comprehensive knowledge of the technical condition of the network and illegal (unidentified) actions of users, enabling to achieve much greater modeling of stormwater inflow to the sewer system. However, they require conducting appropriate optimization studies enabling to calibrate the modeling function parameters. Out of the analyzed models, the best fit of daily sewage volume influent to the WWTP to the daily rainfall depth was achieved for the ARIMAX model.

Funding This work was supported by the statutory activity (FN-70/IŚ/2020) of the Department of Water Supply and Wastewater Disposal, Faculty of Environmental Engineering, Lublin University of Technology, Poland.

Declarations

Conflict of interest Edward Kozłowski declares that he has no conflict of interest. Dariusz Kowalski declares that he has no conflict of interest. Beata Kowalska declares that she has no conflict of interest. Dariusz Mazurkiewicz declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Kozłowski E, Kowalska B, Kowalski D, Mazurkiewicz D. Water demand forecasting by trend and harmonic analysis. *J ACME*. 2018. <https://doi.org/10.1016/j.acme.2017.05.006>.
2. Flaga A, Bosak G, Pistol A, Flaga Ł. Wind tunnel model tests of snow precipitation and redistribution on rooftops, terraces and in the vicinity of high-rise buildings. *J ACME*. 2019. <https://doi.org/10.1016/j.acme.2019.07.007>.
3. Karimi HS, Natarajan B, Ramsey ChL, Henson J, Joshua L, Tedder JL, Kemper E. Comparison of learning-based wastewater flow prediction methodologies for smart sewer management. *J Hydrol*. 2019;577:123977–4012.
4. Garofalo G, Giordano A, Piro P, Spezzano G, Vinci A. A distributed real-time approach for mitigating CSO and flooding in urban drainage systems. *J Netw Comp Appl*. 2017;78:30–42.
5. Garofalo G, Giordano A, Piro P, Spezzano G, Vinci A. A distributed real-time approach for mitigating CSO and flooding in urban drainage systems. *J Netw Comput Appl*. 2017;78:30–42.
6. Piro P, Carbone M, Garofalo G. Distributed vs. concentrated storage options for controlling CSO volumes and pollutant loads. *J Water Pract Technol*. 2010. <https://doi.org/10.2166/wpt.2010.071>
7. Piro P, Carbone M. A modelling approach to assessing variations of total suspended solids (TSS) mass fluxes during storm events. *Hydrol Process*. 2014;28(4):2419–26.
8. Capodaglio AG. Evaluation of modeling techniques for wastewater treatment plant automation. *Water Sci Technol*. 1994;30(2):149–56.
9. Chen J, Ganigué R, Liu Y, Yuan Z. Real-time multistep prediction of sewer flow for online chemical dosing control. *J Environ Eng*. 2014;140:04014037.

10. Liu Y, Ganigué R, Sharma K, Yuan Z. Event-driven model predictive control of sewage pumping stations for sulfide mitigation in sewer networks. *Water Res.* 2016;98:376–83.
11. Hernes RR, Gagne AS, Abdalla EMH, Braskerud BC, Alfredsen K, Muthanna TM. Assessing the effects of four SUDS scenarios on combined sewer overflows in Oslo, Norway: evaluating the low impact development module of the mike urban model. *Hydrol Res.* 2020 (**in press**).
12. Rossman LA, Supply W. Storm water management model, quality assurance report: dynamic wave flow routing. US Environmental Protection Agency, Office of Research and Development, National Research Management Research Laboratory; 2006.
13. Dirckx G, Schütze M, Kroll S, Thoeys C, De Gueudre G, Van De Steene B. RTC versus static solutions to mitigate csos impact. In: 12nd international conference on urban drainage, Porto Alegre, Brazil; 2011.
14. Achleitner S, Möderl M, Rauch W. City drain©—an open source approach for simulation of integrated urban drainage systems. *Environ Model Softw.* 2007;22(8):1184–95.
15. Pleau M, Colas H, Lavalley P, Pelletier G, Bonin R. Global optimal real-time control of the Quebec urban drainage system. *Environ Model Softw.* 2005;20:401–13.
16. Fu G, Butler D, Khu ST. Multiple objective optimal control of integrated urban wastewater systems. *Environ Model Softw.* 2008;23(2):225–34.
17. Schütze M, Campisano A, Colas H, Schilling W, Vanrolleghem PA. Real time control of urban wastewater systems: where do we stand today? *J Hydrol.* 2004;299(3):335–48.
18. Beeneken T, Erbe V, Messmer A, Reder C, Rohlfing R, Scheer M, Schuetze M, Schumacher B, Weilandt M, Weyand M. Real time control (RTC) of urban drainage systems—a discussion of the additional efforts compared to conventionally operated systems. *Urban Water J.* 2013;10(5):293–9.
19. Darsono S, Labadie JW. Neural-optimal control algorithm for real-time regulation of in-line storage in combined sewer systems. *Environ Model Softw.* 2007;22:1349–61.
20. Vezzaro L, Grum M. A generalised dynamic overflow risk assessment (DORA) for real time control of urban drainage systems. *J Hydrol.* 2014;515:292–303.
21. Li J, Sharma K, Liu Y, Jiang G, Yuan Z. Real-time prediction of rain-impacted sewage flow for on-line control of chemical dosing in sewers. *Water Res.* 2019;149:311–21.
22. Zhang D, Martinez N, Lindholm G, Ratnaweera H. Manage sewer in-line storage control using hydraulic model and recurrent neural network. *Water Resour Manag.* 2018;32:2079–98.
23. Jean MÈ, Duchesne S, Pelletier G, Pleau M. Selection of rainfall information as input data for the design of combined sewer overflow solutions. *J Hydrol.* 2018;565:559–69.
24. Carstensen J, Nielsen MK, Strandbaek H. Prediction of hydraulic load for urban storm control of a municipal WWT plant. *Water Sci Technol.* 1998;37(12):363–70.
25. El-Din AG, Smith DW. A neural network model to predict the wastewater inflow incorporating rainfall events. *Water Res.* 2002. [https://doi.org/10.1016/s0043-1354\(01\)00287-1](https://doi.org/10.1016/s0043-1354(01)00287-1).
26. Wei X, Kusiak A, Sadat HR. Prediction of influent flow rate: data-mining approach. *J Energy Eng.* 2012;139:118–23.
27. Zhang D, Lindholm G, Ratnaweera H. Use long short-term memory to enhance internet of things for combined sewer overflow monitoring. *J Hydrol.* 2018;556:409–18.
28. Koronacki J, Mielniczuk J. Statistics for students of technical and natural sciences (Statystyka dla studentów kierunków technicznych i przyrodniczych), Wydawnictwa Naukowo-Techniczne (in Polish); 2009.
29. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2014. <https://otexts.com/fpp2/>
30. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2018.
31. Shumway RH, Stoffer DS. Time series analysis and its applications: with r examples. Berlin: Springer; 2017.
32. Wayne HLG, Woodward A, Elliott AC. Applied time series analysis with r. Milton Park: Taylor & Francis Inc.; 2017.
33. Brown RG. Smoothing, forecasting and prediction of discrete time series. North Chelmsford: Courier Corporation; 2004.
34. Kosicka E, Kozłowski E, Mazurkiewicz D. The use of stationary tests for analysis of monitored residual processes. *Eksploatacja i Niezawodność Maint Reliab.* 2015;17(4):604–9. <https://doi.org/10.17531/ein.2015.4.17>.
35. Kozłowski E. Time series analysis and identification (Analiza i identyfikacja szeregów czasowych), Politechnika Lubelska (in Polish); 2015.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations