



Wastewater inflow time series forecasting at low temporal resolution using SARIMA model: a case study in South Australia

Phuong Do¹ · Christopher W. K. Chow^{1,2} · Raufdeen Rameezdeen¹ · Nima Gorjian^{1,3}

Received: 20 January 2022 / Accepted: 6 May 2022 / Published online: 20 May 2022
© The Author(s) 2022

Abstract

Forecasts of wastewater inflow are considered as a significant component to support the development of a real-time control (RTC) system for a wastewater pumping network and to achieve optimal operations. This paper aims to investigate patterns of the wastewater inflow behaviour and develop a seasonal autoregressive integrated moving average (SARIMA) forecasting model at low temporal resolution (hourly) for a short-term period of 7 days for a real network in South Australia, the Murray Bridge wastewater network/wastewater treatment plant (WWTP). Historical wastewater inflow data collected for a 32-month period (May 2016 to December 2018) was pre-processed (transformed into an hourly dataset) and then separated into two parts for training (80%) and testing (20%). Results reveal that there is seasonality presence in the wastewater inflow time series data, as it is heavily dependent on time of the day and day of the week. Besides, the SARIMA (1,0,3)(2,1,2)₂₄ was found as the best model to predict wastewater inflow and its forecasting accuracy was determined based on the evaluation criteria including the root mean square error (RMSE = 5.508), the mean absolute value percent error (MAPE = 20.78%) and the coefficient of determination ($R^2 = 0.773$). From the results, this model can provide wastewater operators curial information that supports decision making more effectively for their daily tasks on operating their systems in real-time.

Keywords Wastewater inflow · Forecasting · Time series modelling · SARIMA model

Introduction

The operations of wastewater pumping systems or networks consume a tremendous amount of electrical energy to transfer sewage and with both the financial and energy inefficiency issues which can be handled by improving management practices (Galve et al. 2021; Mirra et al. 2020). Meanwhile, practical guidance for pumping operations is generally not available; thus, wastewater operators activate or deactivate the pumps only according to their own expert

knowledge and experience of the system generally resulting in higher operating energy costs (Kim et al. 2006). A pump switching program that properly controls pump on/off applied into the wastewater network can lead to a great reduction in energy costs (Wei et al. 2013), especially when the pumps are planned to operate with the precise estimation of electricity spot market prices and wastewater inflow rates (Do et al. 2022, 2021). Wastewater inflow forecasting plays an essential role in controlling pumping system of a wastewater network (Piri et al. 2021). The quantity of incoming wastewater to the network/wastewater treatment plant (WWTP) can be used to pre-schedule pump operations. Therefore, to achieve optimal schedules for wastewater pumps, it is best to forecast influent flow rate in advance as one of the significant parameters (Kim et al. 2016; Zeng et al. 2016; Wei and Kusiak 2015).

In the recent literature, there have been studies focused on projections of wastewater inflow rate to the WWTPs using different data-driven approaches which can be separated into three categories. The first one is the machine learning (ML) method. Wei et al. (2013) applied four ML algorithms, including multilayer perceptron neural network

Responsible Editor: Marcus Schulz

✉ Christopher W. K. Chow
christopher.chow@unisa.edu.au

¹ Sustainable Infrastructure and Resource Management (SIRM), UniSA STEM, University of South Australia, Mawson Lakes, Adelaide, SA 5095, Australia

² Future Industries Institute, University of South Australia, Adelaide, SA 5095, Australia

³ South Australian Water Corporation, Adelaide, South Australia, Australia

(MLPNN), random forest, boosted tree, and support vector machine to model the quantity of influent flow. The MLPNN was determined as the best-performing algorithm and therefore chosen to produce forecasts. In addition, other ML techniques were also used to predict wastewater inflow rate such as chaos neural network (Li et al. 2007), k-nearest neighbour (Kim et al. 2016) and deep learning (Oliveira et al. 2020). The second data-driven method is the hybrid technique such as adaptive neural fuzzy interference–grey wolf optimiser (ANFIS-GWO) (Dehghani et al. 2019) and multimodal and ensemble-based deep learning (ME-DeepL) (Heo et al. 2021). The last one is the conventional data-driven method such as the autoregressive integrated moving average (ARIMA) model. ARIMA is developed with a time series which is a set of data acquired at evenly spaced time intervals; therefore, it is also called time series model. It has been proven as an effective method in constructing forecasting models for wastewater inflow to WWTPs. Kim et al. (2006) anticipated daily influent rate and properties by developing an ARIMA model based on daily data collected for 150 days. Research outcomes showed good forecast results for 1–7 days ahead. Nevertheless, to enhance the reliability of the proposed forecasting model, a sufficient data quantity was required as the collected datasets did not exhibit seasonal and annual patterns. ARIMA models were also able to describe weekly (Abunama and Othman 2017) and daily (Boyd et al. 2019) observed and future behaviour of wastewater inflow rate to produce forecasts for case study WWTPs. A comparison study was implemented by Zhang et al. (2019) on forecasting ability of the ARIMA and MLPNN models. The ARIMA model was developed using the wastewater inflow data only, while the MLPNN model included exogenous meteorological variables (e.g. temperature, precipitation). The results indicated reliable daily predictions could be obtained by both models. However, the ARIMA model was proven to have higher accuracy in terms of statistical metrics.

Predicting wastewater flow into the WWTPs is a challenging task. According to Zhang et al. (2019), engineers and operators have to cope with a number of uncertainties and complexities, such as the difficulties in simulating influencing factors on wastewater inflow (e.g. rainfall, runoff and infiltration) and the changes in infrastructure due to aging conditions. Time series models such as ARIMA and its derivatives (e.g. seasonal autoregressive integrated moving average (SARIMA), a model is formed by adding seasonal terms to the ARIMA model to deal with seasonal elements in the data series) can overcome these problems (Zhang et al. 2019). The dynamics of the wastewater inflow rate is expected to follow a certain pattern such as time of the day, day of the week, weekly, monthly or quarterly which means

there is a presence of seasonality in the time series. ARIMA model is inadequate for forecasting in this case; therefore, a seasonal ARIMA (SARIMA) approach needs to be applied to develop predictive models (Hyndman and Athanasopoulos 2018) to address the shortcoming of the ARIMA method.

SARIMA technique has been used to build forecasting models in a wide range of scientific disciplines such as hydrology, meteorology, and climatology (Brito et al. 2021; Liu et al. 2021; Ray et al. 2021). However, there is no comprehensive evaluation of the ability, and reported application of the SARIMA model to forecast wastewater inflow has been found. Besides, in the studies of the ARIMA model as mentioned above, researchers only explored predictive ability of this model for high temporal resolution including daily and weekly forecasts. In a real-time control system, wastewater inflow used as smart controller's input to plan pump on/off schedules in advance should be predicted for low temporal resolution. From these knowledge gaps arise the need for further research on the development of a wastewater inflow forecasting model with temporal resolution of 60 min to generate hourly forecasts for pumping control in real-time. Moreover, this need leads to the framing of a research argument on the forecasting performance of the SARIMA model for hourly wastewater inflow.

This study describes the application of the SARIMA model as a predicting approach to address the seasonality in the wastewater inflow time series and forecast future datapoints. The primary purposes of this paper are to characterize wastewater inflow rate and develop a SARIMA model as an inflow forecasting tool for the Murray Bridge WWTP. The accuracy of the proposed model was evaluated based on three statistical indexes, including the root mean square error (RMSE), the mean absolute value percent error (MAPE) and the coefficient of determination (R^2). The main objectives of this study are as follows: (1) identifying and selecting the best SARIMA forecasting model for a real wastewater network/WWTP, (2) generating low temporal resolution (60 min) wastewater inflow forecasts for a short-term period (7 days) and (3) managerial implications regarding the application of hourly inflow predictions in the real-time wastewater pumping control.

This paper proceeds as follows. The methodology describes the methods used for modelling and forecasting low temporal resolution (hourly) wastewater inflow, the research case study, the process of collecting and preparing data, the step-by-step procedure of forecasting model development and criteria to evaluate its accuracy. The findings on wastewater inflow pattern investigation, model development and prediction are provided as results and discussed. Finally, the conclusion gives a summary and highlights the outcomes of the study.

Methodology

SARIMA model

ARIMA time series model (Box et al. 2015) relies on the analysis of historical data to predict future values with an assumption that data patterns in the past can be utilized to predict data in the future. The ARIMA model consists of three components, including (i) autoregression (AR) which describes the correlation between an observation with its own lagged values, (ii) integration (I) which shows the number of times differencing needs to be performed to make the data series stationary, and (iii) moving average (MA) which represents the correlation between observations and residual errors (Wang et al. 2021; Parmezan et al. 2019).

SARIMA is developed by including additional seasonal component to the ARIMA model which handles the seasonality in the time series. SARIMA model, in general, is a combination of the non-seasonal module (p,d,q) and seasonal module (P,D,Q)_s with seven parameters. It is denoted as SARIMA(p, d, q)(P, D, Q)_s (1); where *p* and *P* are the order of non-seasonal and seasonal AR term; *d* and *D* are the degree of non-seasonal and seasonal differencing; *q* and *Q* are the order of non-seasonal and seasonal MA term; and *s* is the length of seasonality in the time series. For example, in an hourly time series, *s* = 24; in a daily time series, *s* = 7; in a monthly time series, *s* = 12; and in a quarterly time series, *s* = 4.

SARIMA, as a member of the ARIMA model family, works best when it is applied for a long and stable time series (Dimri et al. 2020). SARIMA method requires a medium to long length time series that consists of at least 50 data points. It has a strong dependence on the historical data; therefore, the continuity of data is required to be guaranteed (Zhou et al. 2019).

The ARIMA model family such as AR and ARIMA has been a widely used technique for wastewater inflow predictions. However, the SARIMA model, an extended version of the ARIMA model has not yet been applied in the same field. The ARIMA model is utilized if there is no presence of seasonality in a time series. In case a seasonality pattern exists, the SARIMA model needs to be applied (Hyndman and Athanasopoulos 2018). In the time series data, seasonality is observed when the changes in data have a regular pattern that repeats over a certain period. Seasonality is a known and fixed frequency cycle (Hyndman and Athanasopoulos 2018). There are different seasonality types such as time of the day, day of the week, weekly, monthly and quarterly.

Case study

The Murray Bridge wastewater network in South Australia, a realistic wastewater network with real data was selected as the case study to apply the proposed SARIMA model. It serves approximately 14,000 people and covers an area of about 14 km² with different land-use types (e.g. residential, commercial, education and recreational). Details of this network and related studies have been published in Do et al. (2021), Gorjian Jolfaei et al. (2019) and Konetschka et al. (2017). Figure 1 shows the schematic diagram of this case study.

In this study, an assumption has been made that the total wastewater inflow to the Murray Bridge WWTP is considered to be equal to the total flow collected from sources/catchments then transferred by numerous pump stations in the Murray Bridge wastewater network. At the WWTP, the flow meter is installed; therefore, data is available to be used to develop the forecasting model for wastewater inflow.

Data collection and pre-processing

Data preparation was conducted with two stages: (1) collection and (2) pre-processing to gather and transform raw data into a time series dataset used for modelling and forecasting wastewater inflow. The procedures are described in Fig. 2.

- Data collection

The historical wastewater inflow data of the Murray Bridge WWTP for 32 months from 7 May 2016 to 31 December 2018 were sourced from the SA Water's Operational Data Store (ODS). The raw dataset with 1,149,637 data points in total was unevenly spaced which included actual measurements of inflow rate to the WWTP at different sampling times. They were recorded at intervals of 3 s at a minimum and 3 h 50 min at a maximum, and mostly every 5 and 55 s.

- Data pre-processing

Data cleaning was first conducted to identify and handle probable inaccurate or irrelevant data. There were 504 data errors detected, including 484 non-numerical, 16 abnormally large and four negative values. They were determined by sorting the dataset ascending and descending. All of them were eliminated to achieve a more consistent and better accuracy dataset to build a predictive model for wastewater inflow.

The filtered wastewater inflow dataset with 1,149,133 data points remaining after error elimination was converted to an hourly time series dataset by averaging data within each 60 min. After the data conversion process, each day in

Fig. 1 Schematic diagram of the Murray Bridge wastewater network

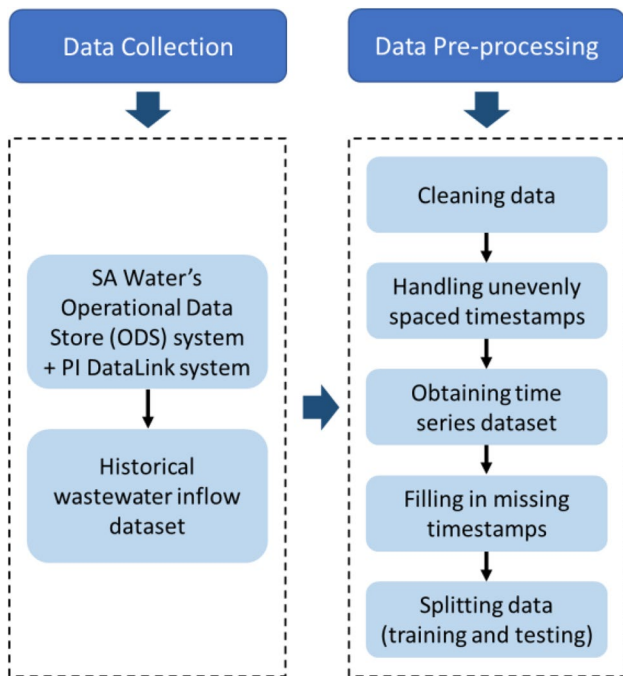
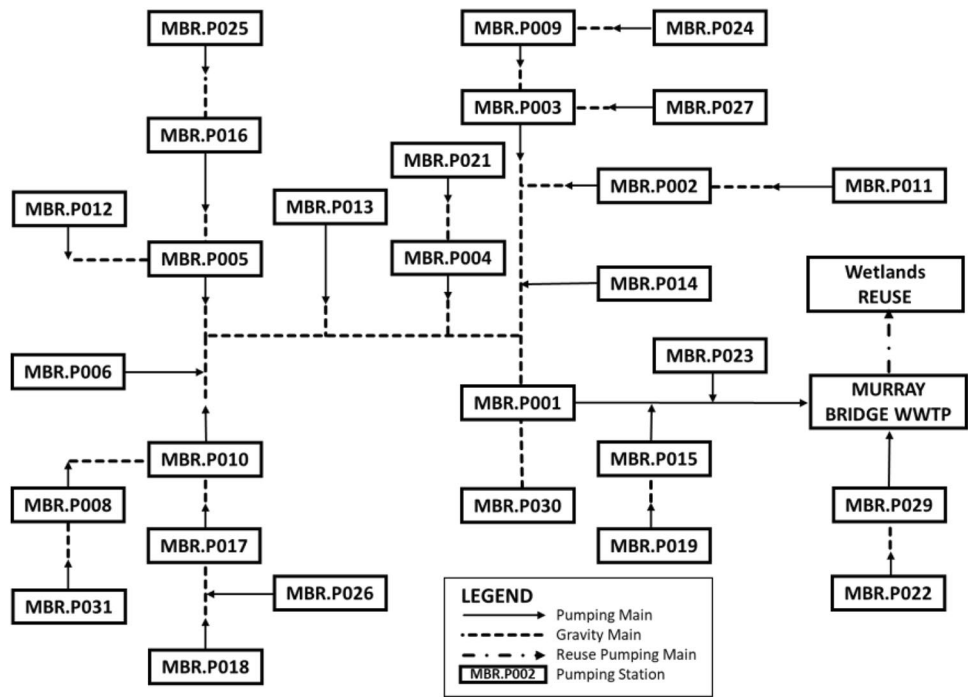


Fig. 2 Data collection and pre-processing procedures

the considered period has 24 records; therefore, an hourly wastewater inflow dataset with 23,256 data points was created. This converted dataset was then inspected to find out any interval without data. Twenty detected missing values accounting for only 0.09% of the entire converted dataset were filled in by averaging the two nearby observations

which are close to the average value of this dataset; therefore, there was no impact on the data.

The hourly wastewater inflow dataset was divided into two parts: training and testing. As stated by Hyndman and Athanasopoulos (2018), typically, the size of the testing set accounts for around 20% of the entire dataset and is ideally at least equal to the longest forecasting duration. Therefore, the ratio of training to testing set is 80:20. The training set that includes data of the first 26 months (May 2016 to June 2018) with 18,840 data points was used for model development. The testing set that consists of data of the remaining 6 months (July to December 2018) with 4416 data points was reserved for model validation. A summary of details of the wastewater inflow datasets is shown in Table 1.

Model development procedure

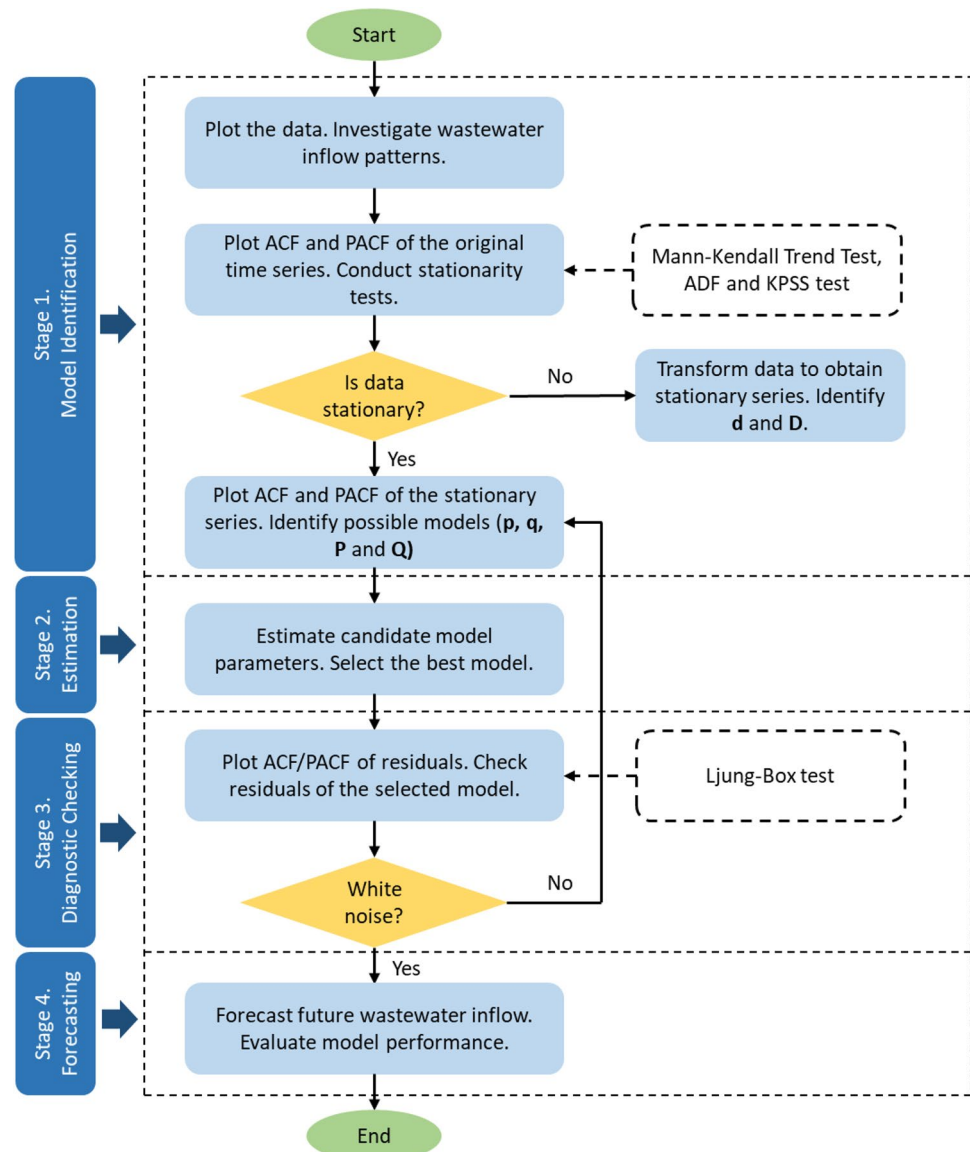
Figure 3 illustrates the flowchart of the step-by-step methodology applied for modelling and forecasting wastewater inflow in this study. The procedures were based on Box and Jenkins methodology (Box et al. 2015) and comprised four stages, including (i) model identification, (ii) parameter estimation, (iii) diagnostic checking and (iv) forecasting. IBM SPSS Statistics 25 was employed as a tool to support the implementation of these four stages.

The procedures of these four stages are described in detail as follows.

- Stage 1. Model identification

Table 1 Wastewater inflow dataset summary

Dataset		Data points	Starting date	Ending date
Raw data		1,149,637	7 May 2016	31 December 2018
Filtered data		1,149,133	7 May 2016	31 December 2018
Converted data (hourly data)	Entire set	23,256	7 May 2016	31 December 2018
	Training set	18,840	7 May 2016	30 June 2018
	Testing set	4416	1 July 2018	31 December 2018

Fig. 3 Flowchart of modelling and forecasting wastewater inflow procedure

The first and most important requirement for the development of SARIMA model is to ensure the wastewater inflow rate time series data is stationary. A time series is considered to be stationary when its statistical features (e.g. mean and variance) are constant over time, or not impacted by time at which the series is observed.

The term “stationarity” is used to imply the stationary status of a time series. In contrast, when a time series exhibits trends (e.g. upward or downward) and/or seasonal patterns (e.g. quarterly, monthly or weekly), it is non-stationary.

On the basis of the above-mentioned requirement, the first step of the model identification stage was checking the stationarity of the data. The training set (see Table 1) was used in this stage and from this section onwards, it is called the original training time series. This wastewater inflow dataset was plotted to provide an initial guess about its stationarity features (Jalil and Rao 2019). Boxplots of wastewater inflow grouped by time of the day and day of the week were used to analyse possible intraday and intraweek patterns visually. Next, to statistically test the data stationarity, the Mann–Kendall Trend Test (Kendall 1975; Mann 1945) was undertaken to examine whether there is an increasing or decreasing trend in the time series. This is a commonly used test for hydro-meteorology time series such as streamflow, rainfall and temperature (Kabbilawsh et al. 2020). The Mann–Kendall test is a non-parametric test that is less impacted by the presence of outliers compared to other parametric tests (Praveen et al. 2020; Wang et al. 2020; Hamed 2009). Additionally, further statistical tests were also implemented to mathematically confirm the stationarity condition of the training wastewater inflow series, including a unit root test, Augmented Dickey–Fuller (ADF) test (Dickey and Fuller 1979) and a stationarity test, Kwiatkowski-Philips-Schmidt-Shin (KPSS) (Kwiatkowski et al. 1992). These three tests were performed at the 5% significance level where the alpha value was 0.05 ($\alpha=0.05$) corresponding to the 95% confidence interval. Two opposing hypotheses were set up for each test, including null hypothesis H_0 and alternative hypothesis H_a . The purpose of hypothesis tests was to decide between H_0 and H_a with rules applied for rejecting the null hypothesis H_0 . Table 2 summarizes statements of the null and alternative hypotheses and decision rules applied for the trend, unit root and stationarity tests.

After checking the stationarity of the training time series using statistical tests, the non-seasonal differencing d and seasonal differencing D were determined. If the series is stationary, it is not required to execute the process of differencing, and the value of parameters d and D is zero. In case the series is non-stationary with the presence of seasonality and trend, the seasonal difference is applied. When there is no trend and seasonality component, the series is transformed by the non-seasonal difference. The value of parameters d and D implies the number of times the wastewater inflow

series needs to be differenced to satisfy stationarity. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the original training time series are created if required to further confirm its stationarity. In this study, the ACF plots depict the correlation coefficient between the wastewater inflow time series and its own lagged values, and the PACF plots measure the partial correlation coefficient between this data series and lagged versions of itself.

The next step was to plot the ACF and PACF of the stationary time series. It could be the original training time series with stationary status or the differenced series after differencing process obtained from the previous step. The non-seasonal and seasonal orders of AR (parameters p and P) and MA (parameters q and Q) were identified based on the ACF and PACF plots. Different values of those parameters were combined to identify possible configurations of (p,d,q) and (P,D,Q) for potential SARIMA models.

- Stage 2. Parameter estimation

In this stage, various potential models identified in stage 1 were examined. The coefficient of determination (R^2), root mean square error (RMSE), and normalized Bayesian information criterion (BIC) were used to select one amongst the potential models. The best model with the optimal set of parameters has the highest R^2 , and the least RMSE and normalized BIC.

- Stage 3. Diagnostic checking

The best model selected in stage 2 was tested to determine whether it adequately captured the behaviour of the wastewater inflow data to the Murray Bridge WWTP. The correlograms ACF and PACF of residuals were plotted to check if the residuals followed a white noise process after fitting a SARIMA(p,d,q)(P,D,Q)_s model to the time series. The difference between observed and fitted data is called residuals. The residuals are white noise when they are identically, independently distributed with a zero mean. If at least 95% of all lags lie within the lower and upper confidence limits, it can be concluded that the selected model can be used for the analysis of the wastewater inflow series.

The Ljung-Box Test (Ljung and Box 1978) was also conducted to detect white noise in the residual time

Table 2 Summary of hypothesis testing for stationarity check

Test	Null hypothesis H_0 and alternative hypothesis H_a	Decision rules
Mann–Kendall test	H_0 : There is no trend in the series H_a : There is a trend in the series	If p -value $\leq \alpha=0.05$, H_0 is rejected If p -value $> \alpha=0.05$, H_0 is failed to be rejected
ADF	H_0 : There is a unit root for the series. The series is non-stationary H_a : There is no unit root for the series. The series is stationary	
KPSS	H_0 : The series is stationary H_a : The series is non-stationary	

series. The hypotheses used for the Ljung–Box test include a null hypothesis H_0 that means the residuals are white noise, and an alternative hypothesis H_a that means the residuals are not white noise. They were performed at the 5% significance level ($\alpha=0.05$). If p -value $\leq \alpha=0.05$, H_0 is rejected, while if p -value $> \alpha=0.05$, H_0 failed to be rejected, and H_a is accepted.

- Stage 4. Forecasting

A model with the highest accuracy in simulating wastewater inflow would be employed to forecast data. Applying the selected SARIMA model, the wastewater inflow series were forecasted using the SPSS software. The predicted values were then matched against the testing set.

Model performance evaluation

In order to determine the precision of the SARIMA model in wastewater inflow predictions, the root mean square error (RMSE), the mean absolute value percent error (MAPE) and the coefficient of determination (R^2) were used as statistical

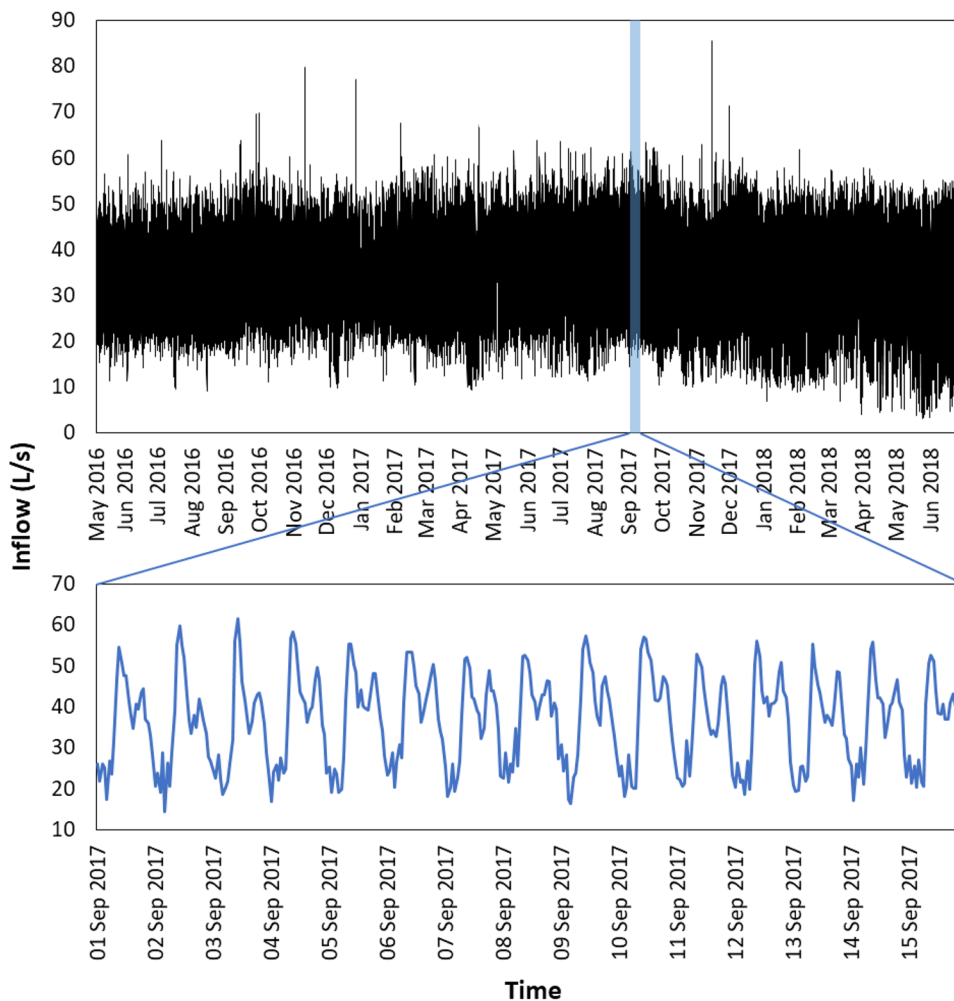
indicators to evaluate the fit of the forecasted to the observed values. Lower values of RMSE and MAPE and a higher value of R^2 imply a more reliable and robust model (Ansari et al. 2018).

Results and discussion

Visualization of the data patterns

The original training time series was used as input for the process of modelling and forecasting hourly wastewater inflow to the WWTP. A plot of this dataset (May 2016 to June 2018) is generated as in Fig. 4. Using this plot, trend and seasonality of the series could be visually identified. A random zoom for the 1–15 September 2017 period was provided to achieve a better insight into the dynamics of the hourly wastewater inflow rates. From this zoom, it could be preliminarily determined that the wastewater inflow rate had no trend and tended to be very low from the beginning of each day, then reached a peak twice during the day. This

Fig. 4 Hourly wastewater inflow to the Murray Bridge WWTP time series plot of the training set (7 May 2016–30 June 2018) and a random zoom (1–15 September 2017)



indicates possible intraday patterns in the time series. Further investigation on trend and seasonal of the series in terms of time of the day and day of the week will be presented in the next section.

Wastewater inflow patterns

Time of the day patterns of the wastewater inflow to the Murray Bridge WWTP are revealed in Fig. 5 in the form of boxplots. The inflow rates were low after midnight till 5:00. The higher wastewater inflow occurred in the early morning, late afternoon and early evening. In particular, from 6:00, it increased then peaked at 10:00. It can be said that hours during the day have a strong influence on the daily high and low wastewater inflow rates. This implies the existence of the intraday seasonality in the wastewater inflow dataset.

Boxplots grouped by day of the week of wastewater inflow data are shown in Fig. 6a. Mondays were often the days with the highest inflow to the network/WWTP which was just slightly greater than that of other weekdays (Tuesdays to Fridays). The inflow rate on Saturdays and Sundays was slightly lower than the remainders of the week. The lower rate on weekends compared to weekdays by time of the day is also shown in Fig. 6b. At every hour of the day, excluding 6 h from 09:00 to 15:00, the weekday inflow rate was higher than that of the weekend. The difference between the higher and lower levels of wastewater inflow by weekdays and weekends indicates it is dependent on the day of the week. Therefore, there is a presence of intraweek seasonality in the wastewater inflow data series.

With the identified intraday and intraweek seasonality components, it can be stated that the wastewater inflow series is non-stationary. To statistically claim the presence

of stationarity in the time series, several trend and stationarity tests were required to implement.

Data stationarity tests

Stationary is a compulsory condition of the data time series to be used for a SARIMA model. If the series is still non-stationary even after certain times of differencing, it is failed to apply the model (Zhang et al. 2019). Before employing the SARIMA technique to develop a forecasting model, the time series data needs to be in a stationary condition. Therefore, the stationarity of the original training time series of hourly wastewater inflow was investigated. The Mann–Kendall trend test, the ADF and KPSS tests, and the ACF and PACF plots were used to verify the data’s stationarity. The results of these statistical tests can be seen in Table 3.

For the Mann–Kendall Trend Test, the calculated p -value (0.16) was greater than the significance level $\alpha=0.05$ indicating that the null hypothesis H_0 failed to be rejected. The result shows there is no downward or upward trend, and the time series is stationary. The ADF test showed the same outcome as the Mann–Kendall trend test. With the p -value of <0.0001 lower than 0.05, the non-stationary null hypothesis was rejected. This confirms there is no unit root in the wastewater inflow series; therefore, the series is stationary. However, the KPSS test indicated a contrary outcome to the other two tests. The calculated p -value (<0.001) was smaller than the significance level $\alpha=0.05$. Thus, the null hypothesis H_0 was rejected which means the wastewater inflow series is concluded to be non-stationary. This may be caused by the strong seasonality of the series as analysed in the previous sections.

The disagreement between results of the KPSS and other tests can be solved by examining the ACF and PACF

Fig. 5 Wastewater inflow pattern by time of the day (7 May 2016–31 December 2018)

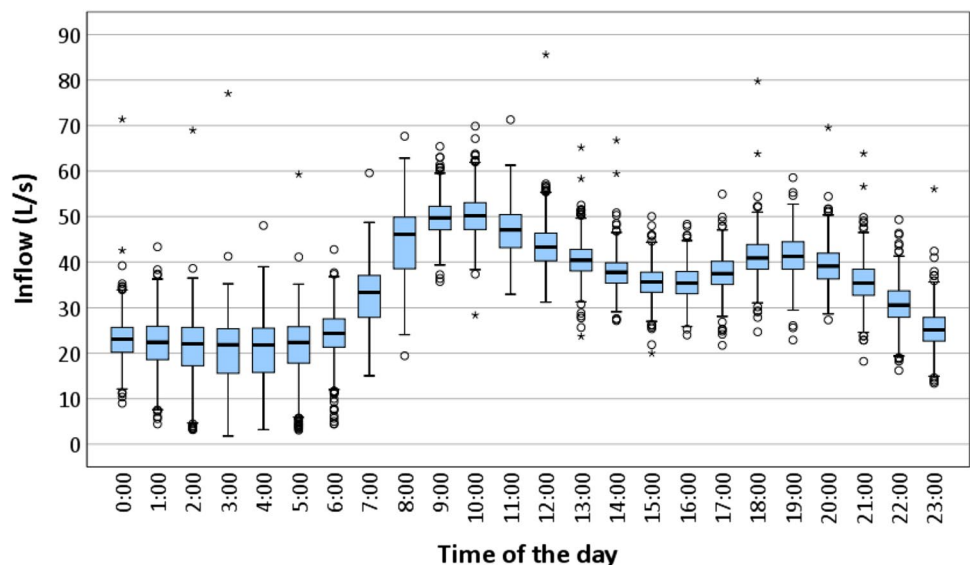


Fig. 6 (a) Wastewater inflow patterns by day of the week and (b) hourly average wastewater inflow by time of the day and by weekday and weekend (7 May 2016–31 December 2018)

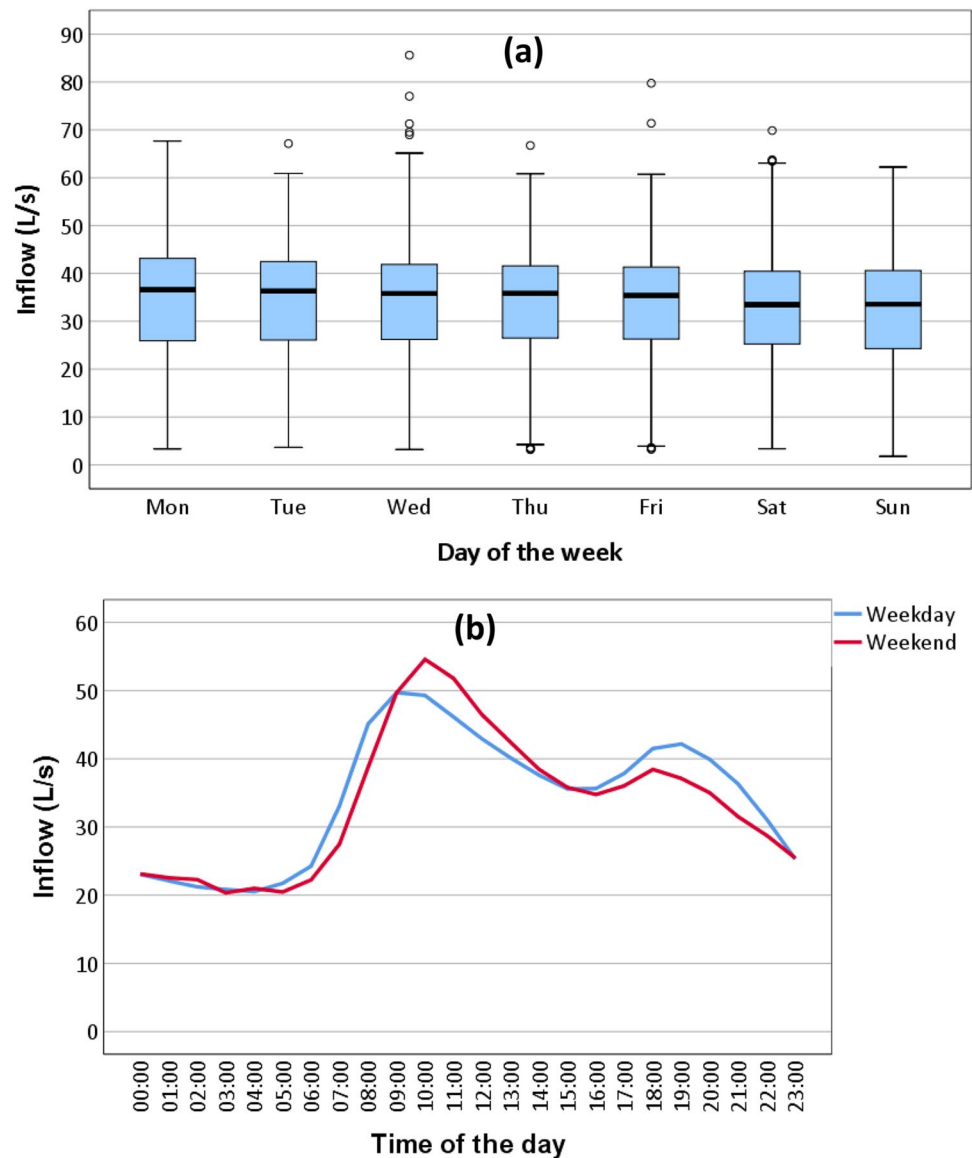


Table 3 Results of the stationarity tests of the original training data series

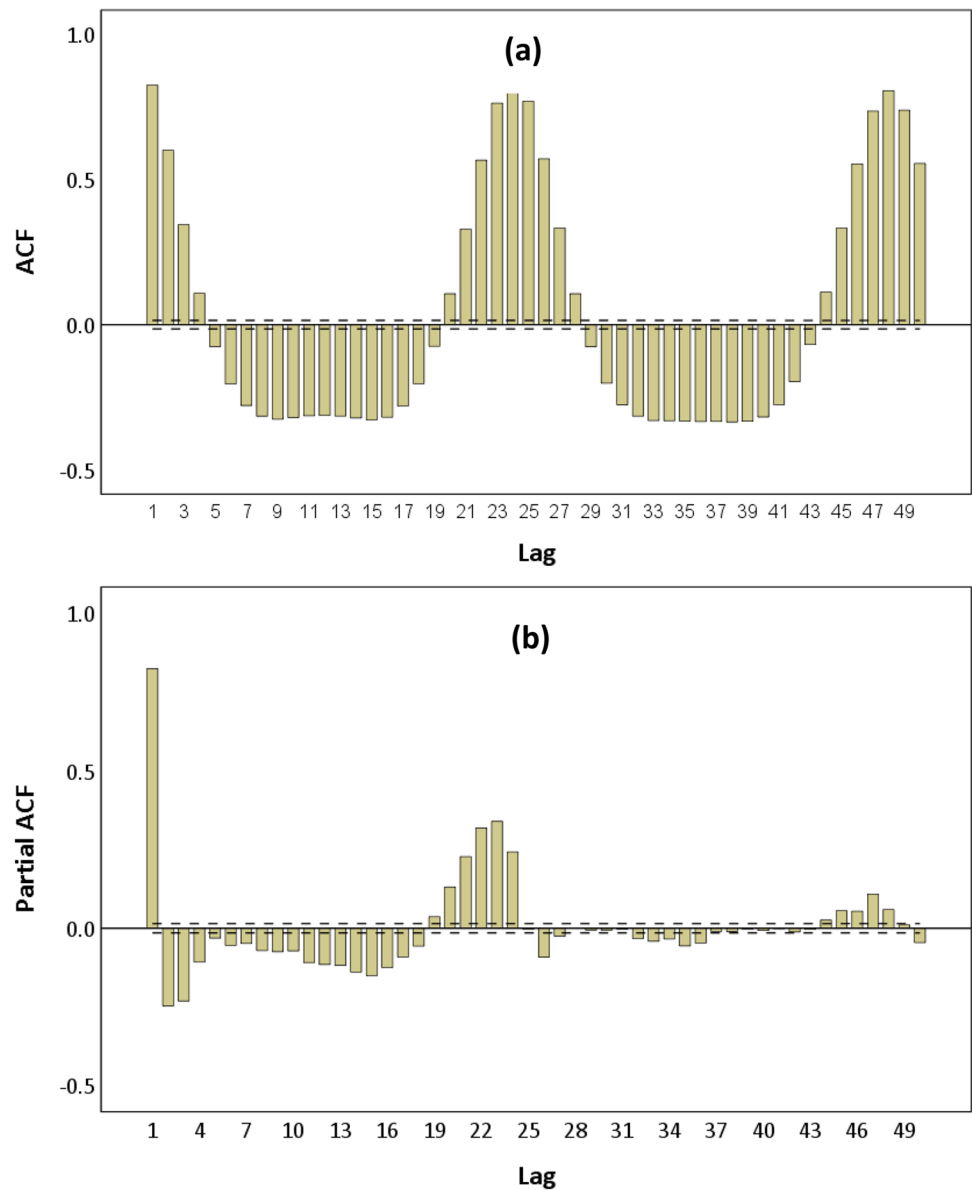
Parameter	Mann–Kendall	ADF	KPSS
<i>p</i> -value	0.160*	<0.0001**	<0.0001**
α	0.05	0.05	0.05
Reject null hypothesis H_0	No	Yes	Yes
Stationary series	Yes	Yes	No

*One-tailed; **two-tailed.

coefficients of the original training time series (Kabbilawsh, Sathish Kumar & Chithra 2020). They are calculated by SPSS and plotted as in Fig. 7a and b. The black dashed lines in each ACF and PACF plot represent the 95% confidence level. The first 50 lags were analysed.

From the ACF plot (Fig. 7a), there are significant peaks at 24 lags such as lag 24 and lag 48 which shows a strong seasonality repeating every 24 time points or 24 h in a day. It can be said that the wastewater inflow data is seasonal with period of seasonality $s=24$. The ACF coefficients move in a sinusoidal wave pattern that is clear evidence of the presence of seasonality that makes the original training time series non-stationary. The existence of seasonality (or seasonal components) in a time series can be subtracted by seasonal differencing technique (Mills 2019; Brockwell and Davis 2016). Therefore, the first order seasonal differencing $D=1$ and periodicity $s=24$ were performed to convert the original training time series to the stationarity form and satisfy the requirement of SARIMA modelling. Figure 8 shows the transformed hourly wastewater inflow training time series and a zoom for the first 15 days of September 2017. The

Fig. 7 (a) ACF and (b) PACF plots of the original training wastewater inflow time series



fluctuations of this series at zero and constant mean demonstrate that it is stationary.

The first order seasonal differenced series obtained by the transformation process was then verified for stationarity using three statistical tests, including the Mann–Kendall Trend Test, ADF and KPSS tests. Table 4 reports results of these three stationarity tests for the first order seasonal differenced series. The p -value resulting from the Mann–Kendall and KPSS test was greater than $\alpha = 0.05$ which means it was failed to reject the corresponding null hypotheses. For the ADF test, the null hypothesis was rejected. All these results infer that the first order seasonal differenced series is stationary and can be used for the SARIMA application.

During the process of analysing and converting the non-stationary wastewater inflow series into stationary, the

non-seasonal differencing was not required, so the value of parameter d is zero. With the seasonal difference $D = 1$ and period of seasonality $s = 24$ as identified previously, SARIMA($p,0,q$)($P,1,Q$)₂₄ models were suggested for further investigation. In the next section, values of parameters p , q , P and Q will be found.

Model selection

The SARIMA($p,0,q$)($P,1,Q$)₂₄ model were ascertained by potential values for the non-seasonal AR order (p), non-seasonal MA order (q), seasonal AR order (P), and seasonal MA order (Q). ACF and PACF plots of the stationary wastewater inflow series which was seasonally differenced with

Fig. 8 Plot of the first order seasonal differenced hourly wastewater inflow for the training time series ($d=0$, $D=1$, and $s=24$) and a random zoom (1–15 September 2017)

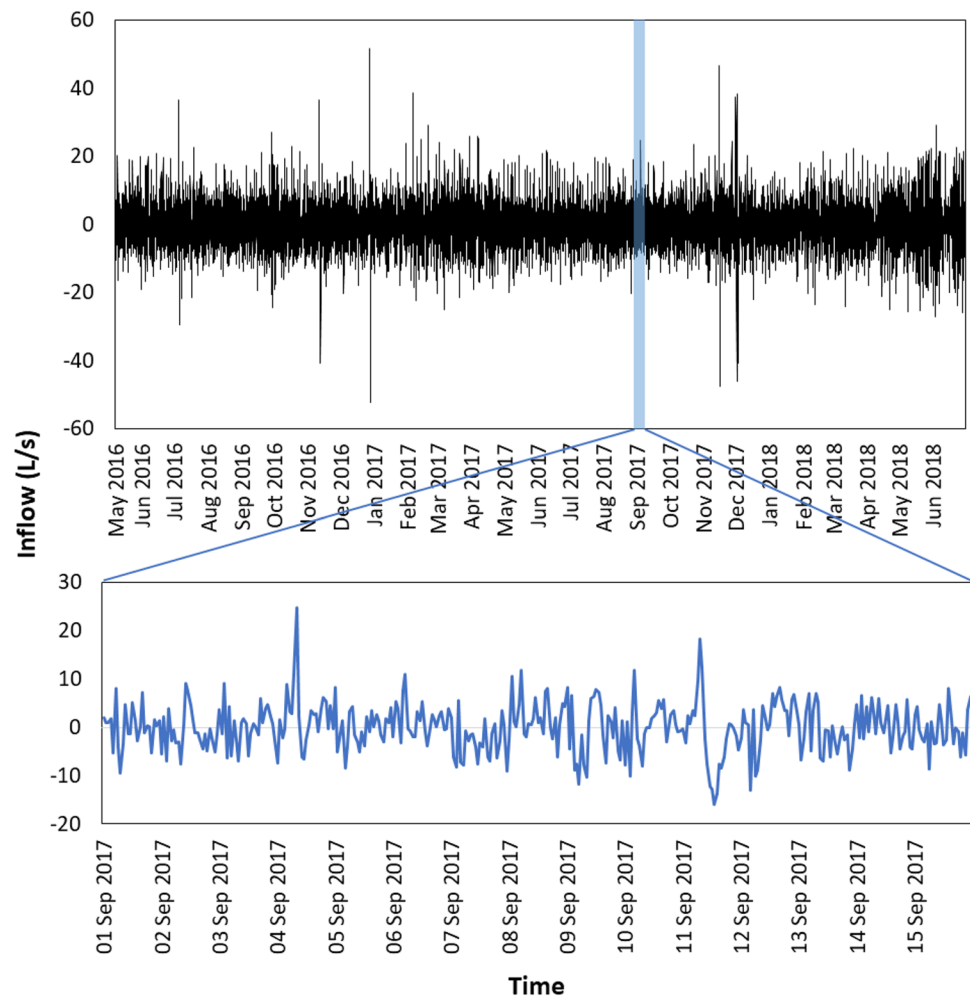


Table 4 Stationarity tests for the first order seasonal differenced series

Parameter	Mann–Kendall	ADF	KPSS
p -value	0.968*	<0.0001**	1.000**
α	0.05	0.05	0.05
Reject null hypothesis H_0	No	Yes	No
Stationary series	Yes	Yes	Yes

*One-tailed; **two-tailed.

$D=1$ and periodicity $s=24$ (see Fig. 9a and b) were used to identify the unknown parameters.

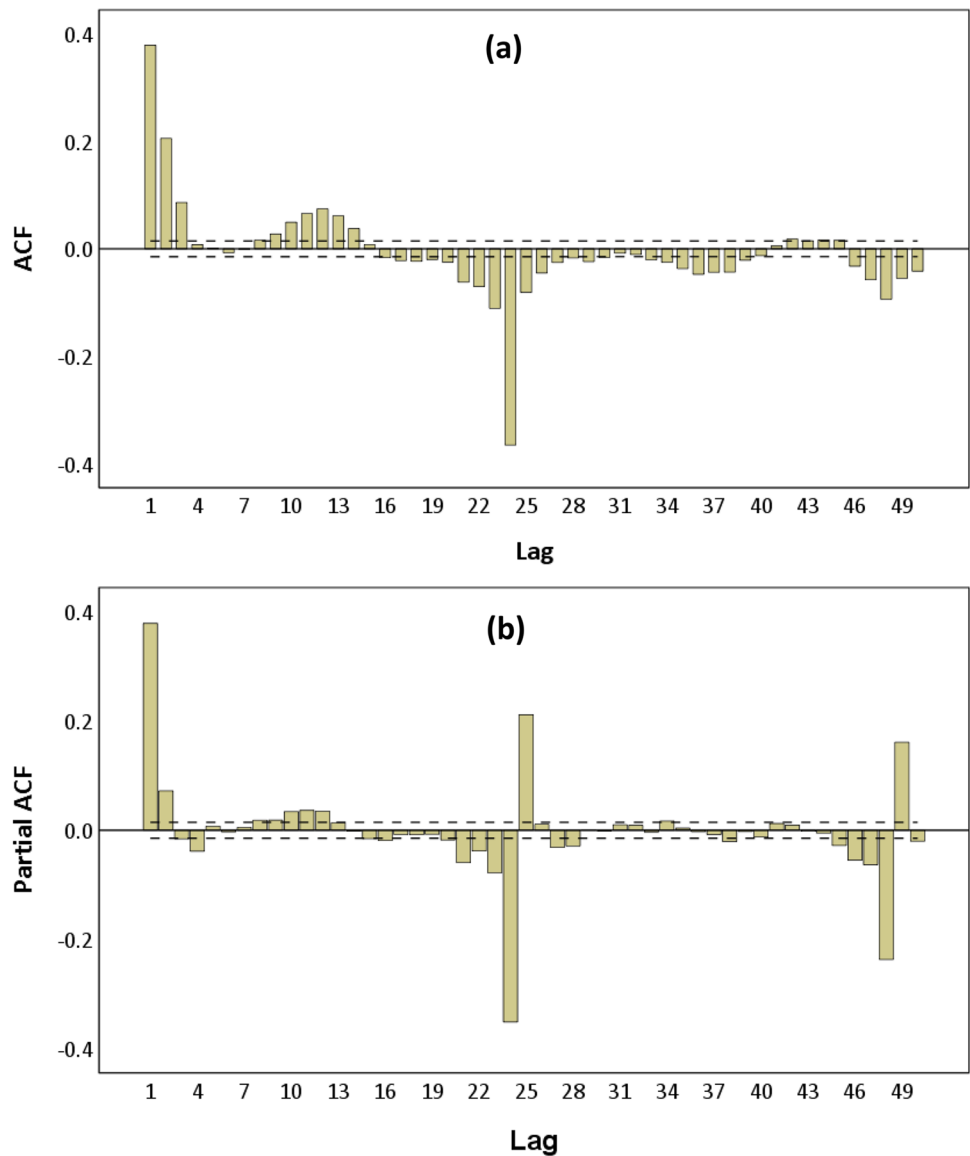
The behaviour of seasonal lags (e.g. lags 24 and 48) and non-seasonal lags which are lags of the first span of periodicity (lags 1–23) in the ACF plot (Fig. 9a) was investigated to determine the parameters q and Q , while in the PACF plot (Fig. 9b), the parameters p and P . From the ACF plot, at lags 1, 2 and 3, significant autocorrelations crossed outside the lower and upper confidence limits that indicate appropriate values of the parameter q .

The continuity of significant autocorrelations disappeared at lags 4–7 as they lied between the lower and upper confidence limits. Thus, the significant autocorrelations at other non-seasonal lags in the first seasonal multiples of 24 were not taken into consideration. There were also significant autocorrelations at lags 12 and 24 of the ACF plot. This means the parameter Q could be 1 or 2. Similarly, significant autocorrelations at lags 1 and 2 observed from the PACF plot imply the potential values of parameter p . Seasonal lags 12 and 24 in the PACF plot with significant autocorrelations indicate 1 and 2 could be the values of the parameter P . With possible values of p , P , q and Q , 24 configurations of those parameters were combined corresponding to 24 potential models.

Determine the optimum parameters

The selection of the best fitting model from 24 potential ones was based on the lowest RMSE and normalized BIC and the highest R^2 . Table 5 presents the results of those evaluation metrics for all potential models.

Fig. 9 (a) ACF and (b) PACF plots of the first seasonally differenced series ($d=0, D=1$ and $s=24$)



Both $SARIMA(2,0,2)(2,1,2)_{24}$ and $SARIMA(1,0,3)(2,1,2)_{24}$ had the smallest values of RMSE (4.113) and normalized BIC (2.833). However, the value of R^2 for $SARIMA(1,0,3)(2,1,2)_{24}$ was higher than that of $SARIMA(2,0,2)(2,1,2)_{24}$ and the highest amongst other models (0.850). It can be concluded that $SARIMA(1,0,3)(2,1,2)_{24}$ is the best model which satisfies the given conditions.

Diagnostic checking

Diagnostic checking was conducted with the purpose of testing the residuals of the best model $SARIMA(1,0,3)(2,1,2)_{24}$ to identify if the SARIMA model sufficiently represents the statistical features of the observed wastewater inflow time

series. Figure 10 shows the ACF and PACF residual plots of the selected model. All residuals lie between 95% confidential limits that indicate there is no autocorrelation amongst residuals; thus, the residuals are white noise.

White noise of the residuals was further tested by the Ljung-Box test to mathematically confirm its absence or existence. This is a diagnostic tool to determine if the residuals of a time series model are independent and identically distributed or if autocorrelation in a time series is different from zero. The results showed that the p -value (0.19) was greater than 0.05 indicating the null hypothesis failed to be rejected. This means the residuals are white noise. In other words, $SARIMA(1,0,3)(2,1,2)_{24}$ removes the residual dependency from the wastewater inflow time series. Therefore, the proposed model passes the required check.

Table 5 SARIMA potential models

	SARIMA model	RMSE	Normalized BIC	R^2
1	(1,0,1)(1,1,1) ₂₄	4.127	2.838	0.848
2	(1,0,2)(1,1,1) ₂₄	4.124	2.837	0.849
3	(1,0,3)(1,1,1) ₂₄	4.122	2.836	0.849
4	(2,0,1)(1,1,1) ₂₄	4.126	2.838	0.849
5	(2,0,2)(1,1,1) ₂₄	4.120	2.835	0.849
6	(2,0,3)(1,1,1) ₂₄	4.122	2.837	0.849
7	(1,0,1)(1,1,2) ₂₄	4.125	2.837	0.849
8	(1,0,2)(1,1,2) ₂₄	4.122	2.836	0.849
9	(1,0,3)(1,1,2) ₂₄	4.120	2.836	0.849
10	(2,0,1)(1,1,2) ₂₄	4.123	2.837	0.849
11	(2,0,2)(1,1,2) ₂₄	4.118	2.835	0.849
12	(2,0,3)(1,1,2) ₂₄	4.120	2.836	0.849
13	(1,0,1)(2,1,1) ₂₄	4.123	2.836	0.849
14	(1,0,2)(2,1,1) ₂₄	4.120	2.835	0.849
15	(1,0,3)(2,1,1) ₂₄	4.118	2.835	0.849
16	(2,0,1)(2,1,1) ₂₄	4.122	2.836	0.849
17	(2,0,2)(2,1,1) ₂₄	4.116	2.834	0.849
18	(2,0,3)(2,1,1) ₂₄	4.118	2.835	0.849
19	(1,0,1)(2,1,2) ₂₄	4.118	2.834	0.849
20	(1,0,2)(2,1,2) ₂₄	4.115	2.833	0.849
21	(1,0,3)(2,1,2) ₂₄	4.113	2.833	0.850
22	(2,0,1)(2,1,2) ₂₄	4.116	2.834	0.849
23	(2,0,2)(2,1,2) ₂₄	4.113	2.833	0.849
24	(2,0,3)(2,1,2) ₂₄	4.131	2.842	0.848

Wastewater inflow forecasting

The ability of the proposed SARIMA model in predicting wastewater inflow data was assessed in this last stage. The testing dataset (1 July to 31 December 2018) was used

for the model validation procedure. The SARIMA(1,0,3)(2,1,2)₂₄ model was directly utilized for the entire testing process. There were no forecasts generated outside the testing period, as this study mainly focuses on the demonstration of the ability of the developed model in predicting future values rather than the actual wastewater inflow rate predictions for the case study WWTP.

The fitness of the observed and forecasted hourly wastewater inflow rate is discussed to determine the quality of the proposed SARIMA model. The mean of the observed and forecasted wastewater inflow rate was 33.26 L/s and 33.35 L/s, respectively. The difference between these two values is only 0.03% that indicates a good fit relationship. In addition, the results of statistical tests were as follows: RMSE = 5.508, MAPE = 20.78% and $R^2 = 0.773$. The RMSE about two times lower than the standard deviation (11.56) is an indication of the good prediction (Boyd et al. 2019). A high-quality forecasting model also has a low value of MAPE. In the previous studies, the results of MAPE were within the range of 71–78% in Zhang et al. (2019) and from 20 to 94% for 4 out of 5 case study WWTPs in Boyd et al. (2019). It can be concluded that a low MAPE value was achieved in this research. Besides, the value of R^2 which is larger than 0.5 also indicate relatively good predictions (Alsharif et al. 2019). As a result, SARIMA(1,0,3)(2,1,2)₂₄ is considered as a reliable forecasting model for wastewater inflow to the Murray Bridge wastewater network/WWTP.

Future forecasts data for the 1-week period (1–7 July 2018) using the proposed SARIMA model are illustrated in Fig. 11. The figure also compares observed and forecasted wastewater inflow, and 95% upper confidential limit (UCL) and lower confidential limit (LCL). SARIMA(1,0,3)(2,1,2)₂₄ in general has the capability to provide future predictions for the wastewater inflow. The forecasted data relatively

Fig. 10 Residuals of ACF and PACF of the SARIMA(1,0,3)(2,1,2)₂₄ model

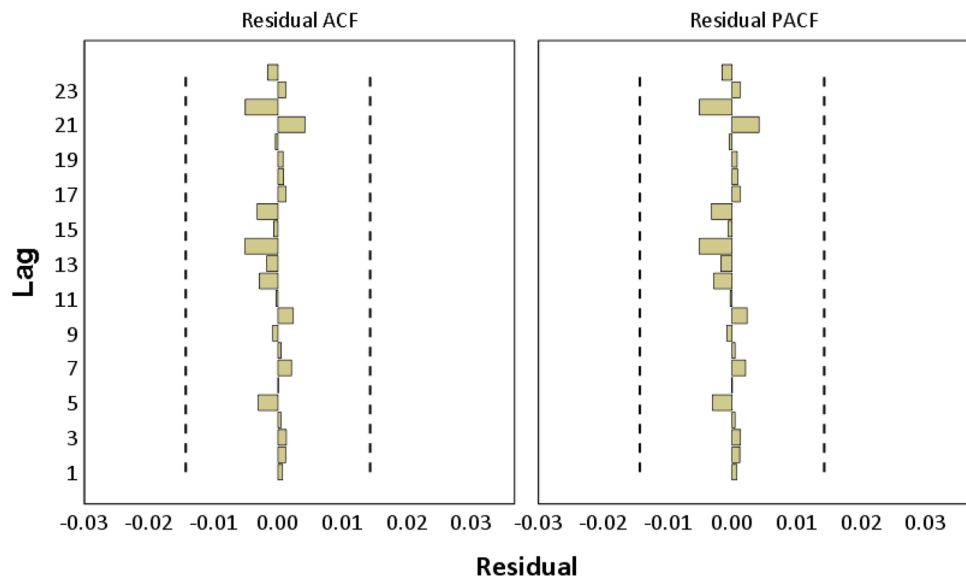
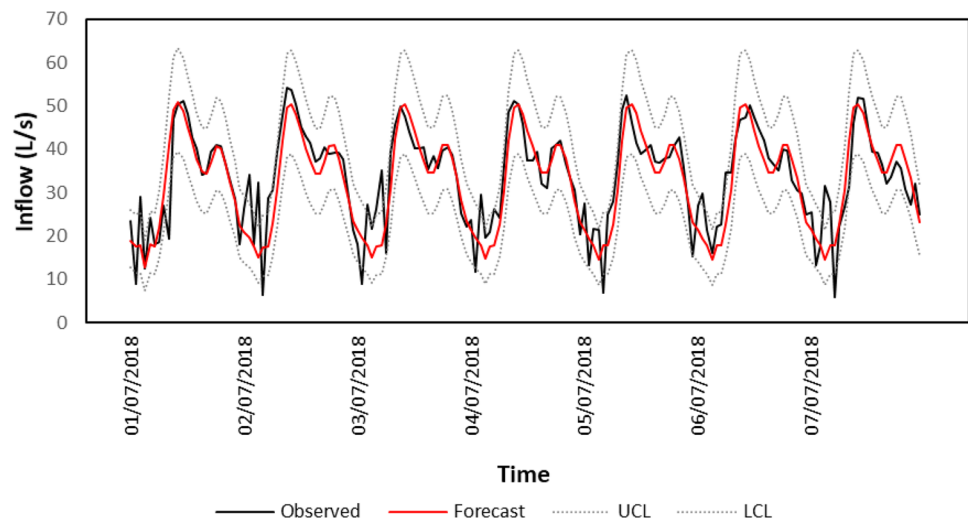


Fig. 11 Observed and forecast wastewater inflow from 1 to 7 July 2018



matched the observed data during morning until midnight from 6 a.m. to 12 a.m. However, it underestimated/overestimated the wastewater inflow during after midnight hours from 1–5 p.m. This could be because ARIMA family models only approximate the data patterns in the past, as the structure of the underlying data mechanism is not explained (YoosefDoost et al. 2017).

The predictions of wastewater inflow to the Murray Bridge wastewater network/WWTP generated based on an hourly time series dataset are useful for the smart wastewater pump controller. This controller operates pumps with consideration of two inputs including wastewater inflow rate and electricity spot price. Details on this smart controller are presented in Do et al. (2021). The forecasts of its two inputs can support the operators in planning the pump schedules during the day in advance. Therefore, wastewater inflow forecasts for low temporal resolution such as 60 min as in this paper are required to accomplish the task. It is impossible with daily and weekly predictions as in previous studies of ARIMA model. According to Dehghani et al. (2019), 7–10 days ahead forecasts offer sufficient time to schedule pumps. Thus, the forecasts every hour for a short-term period of 1 week ahead can significantly contribute to preparing operations plans for the WWTP.

Conclusions

This paper mainly focuses on developing and evaluating the ability of the SARIMA model of predicting wastewater inflow rate to the Murray Bridge wastewater network/WWTP in South Australia. The SARIMA method was applied due to its capability to handle shortcoming of the ARIMA model in dealing with seasonal components in the time series. There has been no evidence of this model application in wastewater inflow rate prediction. Besides,

low temporal resolution forecast of 60 min for wastewater inflow using ARIMA family models has not been demonstrated in the literature. This paper came to fill these gaps of knowledge.

SARIMA technique was successful in wastewater inflow modelling and forecasting for the case study WWTP at low temporal resolution with hourly time series data. SARIMA (1,0,3)(2,1,2)₂₄ was identified as the best model amongst potential ones. The orders (p,d,q) and (P,D,Q) of the proposed SARIMA model were diagnostically checked by performing visualization (ACF and PACF graphs), and statistical test (Ljung-Box test) for the residuals. Short-term forecasts for 1 week ahead were shown for the first 7 days of July 2018. The results indicate the proposed SARIMA model provides high accuracy forecasts based on several evaluation criteria including RMSE, MAPE and R^2 .

The wastewater forecasts for low temporal resolution of 60 min generated from the proposed SARIMA model can be utilized as an input for wastewater pump operations optimization model or pump controller in real-time. Wastewater inflow predictions are an important factor in optimizing the pump operations. With high accuracy forecasts, the pumping system reliability is improved, and pump schedules can be set up appropriately in advance with consideration of the predictions of electricity spot prices to obtain electrical energy cost savings.

An advantage of the SARIMA technique is it only requires historical observations to develop forecasting models, as it relies on the behaviour of past data points to predict future data points. However, it is also a limitation as SARIMA could not include other attributes that have influences on the wastewater inflow rate (e.g. rainfall) as its inputs. These influencing factors should be considered for future research to improve the accuracy of the SARIMA wastewater inflow forecasting model. Moreover, a comparative study on forecasting wastewater inflow rate using SARIMA model and

machine-learning-based techniques such as artificial neural network (ANN), random forest (RF) and k-nearest neighbour (k-NN) is recommended to be conducted to further evaluate the ability of the SARIMA. An additional possible research direction is to further validate the proposed forecasting model for the Murray Bridge WWTP in this study by comparing its performance against that of SARIMA models developed for other WWTPs. The performing ability of each model for each WWTP case study will be assessed and compared based on a number of statistical criteria (e.g. RMSE, MAPE and R^2). Hourly wastewater inflow datasets of the same length of time period should be utilized to generate low temporal resolution forecasting models for those WWTPs to achieve the most accurate results of comparison. Finally, different modelling and forecasting with lower/higher temporal resolutions such as 15 min, 30 min, daily and monthly should be investigated to support the wastewater pumping system and WWTP for different operation purposes.

Acknowledgements This research was financially supported by the University of South Australia (UniSA) University President's Scholarships (UPS) and the South Australian Water Corporation (SA Water). The authors would like to acknowledge the support of SA Water staff, Amanda Mussared, Flavio Bressan and Stephen Nguyen.

Author contribution Conceptualization: Phuong Do, Christopher W. K. Chow, Raufdeen Rameezdeen, Nima Gorjian; methodology: Phuong Do; formal analysis and investigation: Phuong Do; writing — original draft preparation: Phuong Do; writing — review and editing: Christopher W. K. Chow, Raufdeen Rameezdeen, Nima Gorjian; funding acquisition: [nil], resources: Nima Gorjian; supervision: Christopher W. K. Chow, Raufdeen Rameezdeen, Nima Gorjian.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions Phuong Do's scholarship was supported by the University of South Australia (UniSA) University President's Scholarships (UPS) and the South Australian Water Corporation (SA Water).

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

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

- Abunama T, Othman F (2017) Time series analysis and forecasting of wastewater inflow into Bandar Tun Razak Sewage Treatment Plant in Selangor, Malaysia. *IOP Conf Ser: Mater Sci Eng* 210:012028. <https://doi.org/10.1088/1757-899X/210/1/012028>
- Alsharif MH, Younes MK, Kim J (2019) Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul. *South Korea Symmetry* 11(2):240. <https://doi.org/10.3390/sym11020240>
- Ansari M, Othman F, Abunama T, El-Shafie A (2018) Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant. *Malaysia Environ Sci Pollut Res* 25(12):12139–12149. <https://doi.org/10.1007/s11356-018-1438-z>
- Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons
- Boyd G, Na D, Li Z, Snowling S, Zhang Q, Zhou P (2019) Influent forecasting for wastewater treatment plants in North America. *Sustainability* 11(6):1764. <https://doi.org/10.3390/su11061764>
- Brito GRA, Villaverde AR, Quan AL, Pérez MER (2021) Comparison between SARIMA and Holt-Winters models for forecasting monthly streamflow in the western region of Cuba. *SN Appl Sci* 3(6):1–12. <https://doi.org/10.1007/s42452-021-04667-5>
- Brockwell PJ, Davis RA (2016) Nonstationary and seasonal time series models. In: *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer, Cham, pp 157–193. https://doi.org/10.1007/978-3-319-29854-2_6
- Dehghani M, Seifi A, Riahi-Madvar, (2019) Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization. *J Hydrol* 576:698–725. <https://doi.org/10.1016/j.jhydrol.2019.06.065>
- Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc* 74(366a):427–431. <https://doi.org/10.1080/01621459.1979.10482531>
- Dimri T, Ahmad S, Sharif M (2020) Time series analysis of climate variables using seasonal ARIMA approach. *J Earth Syst Sci* 129(1):149. <https://doi.org/10.1007/s12040-020-01408-x>
- Do P, Jolfaei NG, Gorjian N, van der Linden L, Ahammed F, Rameezdeen R, Jin B, Chow CWK (2021) Smart scheduling of pump control in wastewater networks based on electricity spot market prices. *Water Conserv Sci Eng* 6(2):79–94. <https://doi.org/10.1007/s41101-021-00104-1>
- Do P, Chow CWK, Rameezdeen R, Gorjian N (2022) Understanding the impact of spot market electricity price on wastewater asset management strategy. *Water Conserv Sci Eng* 7(2):101–117. <https://doi.org/10.1007/s41101-022-00132-5>
- Galve JCA, Sundo MB, Camus DRD, De Padua VMN, Morales RDF (2021) Series type vertical subsurface flow constructed wetlands for dairy farm wastewater treatment. *Civ Eng J* 7(2):292–303. <https://doi.org/10.28991/cej-2020-03091542>
- Gorjian Jolfaei N, Jin B, Chow C, Bressan F, Gorjian N (2019) An optimised energy saving model for pump scheduling in wastewater networks. In: Mathew J, Lim CW, Ma L, Sands D, Cholette ME, Borghesani P (eds) *Asset intelligence through integration and interoperability and contemporary vibration engineering technologies*. Lecture Notes in Mechanical Engineering. Springer, Cham, pp 197–208. https://doi.org/10.1007/978-3-319-95711-1_20

- Hamed K (2009) Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. *J Hydrol* 368(1–4):143–155. <https://doi.org/10.1016/j.jhydrol.2009.01.040>
- Heo S, Nam K, Loy-Benitez J, Yoo C (2021) Data-driven hybrid model for forecasting wastewater influent loads based on multimodal and ensemble deep learning. *IEEE Trans Industr Inform* 17(10):6925–6934. <https://doi.org/10.1109/TII.2020.3039272>
- Hyndman RJ, Athanasopoulos G (2018) *Forecasting: principles and practice*, 2nd edn. OTexts, <<https://otexts.com/fpp2/>>.
- Jalil A, Rao NH (2019) Time series analysis (stationarity, cointegration, and causality). In: Özcan B, Öztürk I (eds) *Environmental Kuznets curve (EKC): a manual*, academic press. London, UK, pp 85–99. <https://doi.org/10.1016/B978-0-12-816797-7.00008-4>
- Kabbilawsh P, Sathish Kumar D, Chithra NR (2020) Trend analysis and SARIMA forecasting of mean maximum and mean minimum monthly temperature for the state of Kerala, India. *Acta Geophys* 68(4):1161–1174. <https://doi.org/10.1007/s11600-020-00462-9>
- Kendall MG (1975) *Rank correlation methods*, 4th edn. Charles Griffin, London
- Kim JR, Ko JH, Im JH, Lee SH, Kim SH, Kim CW, Park TJ (2006) Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model. *Water Sci Technol* 53(4–5):185–192. <https://doi.org/10.2166/wst.2006.123>
- Kim M, Kim Y, Ki H, Piao W, Kim C (2016) Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. *Front Environ Sci Eng* 10(2):299–310. <https://doi.org/10.1007/s11783-015-0825-7>
- Konetschka M, Jin H, Jolfaei NG, Bologianis S, Bressan F, Chow C, Jin B (2017) Developing intelligent, cost saving pump controls for wastewater networks through integration with the electricity spot market. In the proceedings of the OzWater'17 Conference, Sydney, Australia, 16–18 May 2017
- Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *J Econom* 54(1–3):159–178. [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y)
- Li X, Zeng G, Huang G, Li J, Jiang R (2007) Short-term prediction of the influent quantity time series of wastewater treatment plant based on a chaos neural network model. *Front Environ Sci Eng China* 1(3):334–338. <https://doi.org/10.1007/s11783-007-0057-6>
- Liu X, Lin Z, Feng Z (2021) Short-term offshore wind speed forecast by seasonal ARIMA - a comparison against GRU and LSTM. *Energy* 227:120492. <https://doi.org/10.1016/j.energy.2021.120492>
- Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. *Biometrika* 65(2):297–303. <https://doi.org/10.1093/biomet/65.2.297>
- Ma S, Zeng S, Dong X, Chen J, Olsson G (2014) Short-term prediction of influent flow rate and ammonia concentration in municipal wastewater treatment plants. *Front Environ Sci Eng* 8(1):128–136. <https://doi.org/10.1007/s11783-013-0598-9>
- Mann HB (1945) Nonparametric tests against trend. *Econometrica J Econom Soc* 13:245–259. <https://doi.org/10.2307/1907187>
- Mills TC (2019) *Applied time series analysis: a practical guide to modeling and forecasting*. Academic Press, UK
- Mirra R, Ribarov C, Valchev D, Ribarova I (2020) Towards energy efficient onsite wastewater treatment. *Civ Eng J* 6(7):1218–1226. <https://doi.org/10.28991/cej-2020-03091542>
- Oliveira P, Fernandes B, Aguiar F, Pereira MA, Analide C, Novais P (2020) A deep learning approach to forecast the influent flow in wastewater treatment plants. In: Analide C, Novais P, Camacho D, Yin H (eds) *Intelligent data engineering and automated learning – IDEAL 2020*. IDEAL 2020. Lecture Notes in Computer Science, vol 12489. Springer, Cham, pp 362–373. https://doi.org/10.1007/978-3-030-62362-3_32
- Parmezan ARS, Souza VMA, Batista GEAPA (2019) Evaluation of statistical and machine learning models for time series prediction: identifying the state-of-the-art and the best conditions for the use of each model. *Inf Sci* 484:302–337. <https://doi.org/10.1016/j.ins.2019.01.076>
- Piri J, Pirzadeh B, Keshtegar B, Givehchi M (2021) A hybrid statistical regression technical for prediction wastewater inflow. *Comput Electron Agric* 184:106115. <https://doi.org/10.1016/j.compag.2021.106115>
- Praveen B, Talukdar S, Shahfahad MS, Mondal J, Sharma P, Islam ARMT, Rahman A (2020) Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. *Sci Rep* 10(1):10342. <https://doi.org/10.1038/s41598-020-67228-7>
- Ray S, Das SS, Mishra P, Al Khatib AMG (2021) Time series SARIMA modelling and forecasting of monthly rainfall and temperature in the South Asian countries. *Earth Syst Environ* 5(3):531–546. <https://doi.org/10.1007/s41748-021-00205-w>
- Wang F, Shao W, Yu H, Kan G, He X, Zhang D, Ren M, Wang G (2020) Re-evaluation of the power of the Mann-Kendall test for detecting monotonic trends in hydrometeorological time series. *Front Earth Sci* 8:14. <https://doi.org/10.3389/feart.2020.00014>
- Wang X, Tian W, Liao Z (2021) Statistical comparison between SARIMA and ANN's performance for surface water quality time series prediction. *Environ Sci Pollut Res* 28(25):33531–33544. <https://doi.org/10.1007/s11356-021-13086-3>
- Wei X, Kusiak A (2015) Short-term prediction of influent flow in wastewater treatment plant. *Stoch Environ Res Risk Assess* 29(1):241–249. <https://doi.org/10.1007/s00477-014-0889-0>
- Wei X, Kusia A, Sadat HR (2013) Prediction of influent flow rate: data-mining approach. *J Energy Eng* 139(2):118–123. [https://doi.org/10.1061/\(ASCE\)EY.1943-7897.0000103](https://doi.org/10.1061/(ASCE)EY.1943-7897.0000103)
- YoosefDoost A, Sadeghian MS, NodeFarahani M, Rasekhi A (2017) Comparison between performance of statistical and low cost ARIMA model with GFDL, CM2.1 and CGM 3 atmosphere-ocean general circulation models in assessment of the effects of climate change on temperature and precipitation in Taleghan Basin. *Am J Water Resour* 5(4):92–99. <https://doi.org/10.12691/ajwr-5-4-1>
- Zeng Y, Zhang Z, Kusiak A, Tang F, Wei X (2016) Optimizing wastewater pumping system with data-driven models and a greedy electromagnetism-like algorithm. *Stoch Environ Res Risk Assess* 30(4):1263–1275. <https://doi.org/10.1007/s00477-015-1115-4>
- Zhang Q, Li Z, Snowling S, Siam A, El-Dakhkhni W (2019) Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. *Water Sci Technol* 80(2):243–253. <https://doi.org/10.2166/wst.2019.263>
- Zhou P, Li Z, Snowling S, Baetz BW, Na D, Boyd G (2019) A random forest model for inflow prediction at wastewater treatment plants. *Stoch Environ Res Risk Assess* 33(10):1781–1792. <https://doi.org/10.1007/s00477-019-01732-9>

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