Article

A Methodology for Forecasting Dissolved Oxygen in Urban Streams

Stephen Stajkowski 1, Mohammad Zeynoddin 2, Hani Farghaly 1, Bahram Gharabaghi 1,* and Hossein Bonakdari 2,*

1 School of Engineering, University of Guelph, Guelph, ON NIG 2W1, Canada; stajkows@uoguelph.ca (S.S.); hfarghaly17@gmail.com (H.F.)
2 Department of Soils and Agri-Food Engineering, Université Laval, Québec, QC G1V0A6, Canada; mohammad.zeynoddin.1@ulaval.ca
* Correspondence: bgharaba@uoguelph.ca (B.G.); hossein.bonakdari@fsaa.ulaval.ca (H.B.); Tel.: +1-519-824-4120 (ext. 58451) (B.G.); +1-418-656-2131 (H.B.); Fax: +1-418-656-3723 (H.B.)

Received: 21 July 2020; Accepted: 13 September 2020; Published: 15 September 2020

Abstract: Real-time monitoring of river water quality is at the forefront of a proactive urban water management strategy to meet the global challenge of vital freshwater resource sustainability. The concentration of dissolved oxygen (DO) is a primary indicator of the health state of the aquatic habitats, and its modeling is crucial for river water quality management. This paper investigates the importance of the choices of different techniques for preprocessing and stochastic modeling for developing a simple and reliable linear stochastic model for forecasting DO in urban rivers. We describe several methods of evaluation, preprocessing, and modeling for the DO parameter time series in the Credit River, Ontario, Canada, to achieve the optimum data preprocessing and input selection techniques and consequently obtain the optimum performance of the stochastic models as an effective river management tool. The Manly normalization and standardization \( (S_{td}) \) methods were chosen for preprocessing the time series. Modeling the preprocessed time series using the stochastic autoregressive integrated moving average (ARIMA) model resulted in very accurate forecasts with a negligible difference from sole normalization and spectral analysis \( (S_f) \) methods.

Keywords: water resources; stochastic; preprocessing; dissolved oxygen; water quality

1. Introduction

River water quality monitoring programs have been established around the globe to help watershed managers better protect the quality of water in their watersheds [1]. The dissolved oxygen (DO) concentration plays a critical role in regulating various biogeochemical processes and biological communities in river ecosystems. Maintaining sufficient levels of DO in water is critical for water quality because oxygen is needed for the survival and preservation of various aquatic species, including fish, amphibians, benthos, bacteria, and aquatic plants. Therefore, DO is used as a health index for water bodies [2].

Several studies have described the importance of both the diurnal and seasonal fluctuations of DO as it relates to fish habitats [3–5]. The occurrence of low DO concentrations, in a normally well-oxygenated river system, can cause mortality in fish and other aquatic life. When DO concentrations are reduced, aquatic species are forced to lower their activity and alter respiration rates, which will delay their development, and can cause reproductive problems and/or deformities [6].

As observed in both the diurnal and the seasonal fluctuations in the water temperature and the DO concentrations, a strong negative correlation commonly exists between DO concentrations and river water temperatures [7]. Refractory carbonaceous biological oxygen demand (CBOD) and the oxidation of ammonia to form nitrite and nitrate are some of the most significant processes leading...
to low dissolved oxygen concentrations in a river [8]. DO concentrations in natural river ecosystems are affected by many complex processes, including, solubility (atmospheric pressure, water temperature, humidity, and cloud cover) and processes that determine river flow turbulence mixing (stream depth, width, velocity, and roughness) to name a few [9,10].

Most recently, many researchers have employed advanced machine learning methods and artificial intelligence (AI) techniques for forecasting streamflow and water quality parameters [11–14]. Sentas et al. [14] compared the ARIMA model with an artificial neural network (ANN) for modeling short daily DO. They claimed that the ARIMA model produced better results in modeling the DO parameters measured at several depths. Heddad and Kisi [15] performed multi-input AI modeling for DO estimation. They used daily water temperature, PH, specific conductance as least square support vector, multivariate adaptive regression splines, and M5 Model Tree (M5T) model inputs and found that these models produce acceptable results in DO modeling. They also reported that the results vary station by station and no superior model can be chosen. Harvey et al. [16] developed a regression model to forecast the daily, weekly, and monthly water temperature and DO. They also claimed that their regression models produced accurate results linking air temperature and DO. Bertone et al. [17] used a multi-input regression model to forecast a seven-days-ahead DO. They achieved a correlation coefficient higher than 0.8 and low errors. They also utilized preprocessing methods to obtain preferable results. Parmar and Bhardwaj [18], by a limited analysis ARIMA model, modeled monthly water quality parameters (DO, PH, temperature) for the Yamuna river. They obtained promising results from the stochastic modeling of these parameters.

Using multi-input models has the advantage of utilizing more historical data to estimate the studied parameter. However, such an approach is accompanied by accumulated errors from the measurements of each of the inputs. Optimized input selection and the complexity of such models is another problem with this approach. The availability of historical data, gaps in the dataset, and the associated cost of collecting this data is another challenge with applying a multi-input modeling approach. Gaps in the data are remedied by removing the missing time intervals [19] or reconstructed by interpolation [20], which becomes a challenge when the gaps are numerous or the timespans are long. A solution to these problems can be achieved by the adoption of a single time series modeling approach, a proper analysis procedure and preprocessing method, careful selection of inputs, or using simpler models.

A universal pursuit for less complex models for DO forecasting led to the application of stochastic modeling techniques [21,22]. One of the most widely used stochastic models in hydrology is the ARIMA model using nonseasonal parameters. The advantage of the stochastic ARIMA model over deterministic models is that it requires less data, and it is well suited for forecasting [22]. However, the successful application of the stochastic ARIMA model for the forecasting of DO is associated with several challenges, the key issues being proper data preprocessing that is addressed in this study.

Therefore, the main objective of this study is to compare preprocessing methods applied to the DO ARIMA models to develop the most accurate while still simple and practical forecast model, which to our knowledge has not been studied in-depth before. The ultimate goal of the study is to develop a user-friendly model for the practitioners to easily assess the health of urban streams and avoid the challenging problems associated with the application of more complex multiparameter models.

2. Materials and Methods

2.1. ARIMA Model

One of the most widely used stochastic models in hydrology is the ARIMA model, which is defined by using nonseasonal parameters. These nonseasonal parameters are $\phi$ and $\theta$ (autoregressive...
and moving average parameter, respectively). The orders of each parameter are represented by \( p \) and \( q \), respectively, with \( d \) denoting the differencing order. The model’s equation is defined as [23]:

\[
(1 - \phi_1 B_{t-1} - \ldots - \phi_p B_{t-p}) (1 - \theta_1 B_{t-1} - \ldots - \theta_q B_{t-q}) \varepsilon(t) = (1 - \theta_1 B_{t-1} - \ldots - \theta_q B_{t-q}) \varepsilon(t) \tag{1}
\]

\((1 - B_{t-1})^d\) is the \( d \)-th nonseasonal differencing, \( B \) is the backshift operator. The striking feature of the ARIMA model is that it uses the inline differencing method to stationarize the time series. After calculating the model parameters, historical values are used to forecast one-step forward values.

Stationarity is defined as the stability of statistical parameters. Each time series is composed of periodic patterns, steps, gradual upward or downward changes, and a stochastic part. The existence of each one of the first three terms (periodic, steps, gradual trend) in a time series alone or simultaneously causes the time series to be nonstationary. Therefore, the series should be made stationary by one of stationarizing methods. By using the \( \text{Diff} \) method, each data is subtracted from the previous data, and so trends and seasonal patterns in the series are eliminated [24]:

\[
\text{Diff} (\text{DO}) = \text{DO} (t + 1) - \text{DO} (t) \tag{2}
\]

Standardization (\( S_{ld} \)) is one of the preprocessing methods used to modify the data distribution scale so that the observational data has a mean of zero and a standard deviation of one. Standardization eliminates the trend and removes the jump in the variance (based on the variance before and after the jump) [25].

\[
S_{ld} = \frac{\text{DO}(t) - \overline{\text{DO}}(t)}{S_d} \tag{3}
\]

Periodic patterns can be presented by sinusoidal functions. Therefore, by transforming the time series to the frequency domain, these patterns can be removed from the DO time series. The Fourier expansion is employed to perform the spectral analysis (\( S_f \)) by which the repetitive patterns in the time series can be formulated and subtracted from the DO time series [26]. The remaining parts are closer to being stationary.

\[
P (t) = \overline{\text{DO}}(t) + \sum_{z=1}^{k} \left[ (\alpha_z \cos(2\pi f_z t) + \beta_z \sin(2\pi f_z t)) + \varepsilon(t) \right], \quad t = 1, 2, 3, \ldots, N \tag{4}
\]

\[
\alpha_z = \frac{2}{N} \left( \sum_{t=1}^{N} \text{DO} \cos(2\pi f_z t) \right), \quad z = 1, 2, 3, \ldots, k \tag{5}
\]

\[
\beta_z = \frac{2}{N} \left( \sum_{t=1}^{N} \text{DO} \sin(2\pi f_z t) \right), \quad z = 1, 2, 3, \ldots, k \tag{6}
\]

\[
f_z = \frac{z}{N} \tag{7}
\]

\( \overline{\text{DO}}(t) \) is the average of DO data points, \( \varepsilon(t) \), are the residuals, \( \alpha_z \), and \( \beta_z \) are the Fourier coefficients, \( f_z \) equals to the \( z \)-th harmony base frequency, and \( k \) is the maximum harmonic equal to \( 2z \) and \( 2z - 1 \) for even and odd data, respectively. By subtracting the periodic term \( P(t) \) from DO series, the \( S_f \) series is achieved:

\[
S_f = \text{DO}(t) - P(t) \tag{8}
\]

The normal distribution is the basic assumption of statistical models. Therefore, the distribution of the series should be evaluated prior to modeling. Many visual and numerical tests are available to assess the distribution of the DO time series. The Jarque–Bera (\( JB \)) test is one of many numerical tests that evaluates the normality based on Kurtosis and Skewness [27]. In the absence of a normal distribution, normalization transformations are used, with the method developed by Manly [28].
This transformation is the developed form of the Box–Cox (BC) [29] transformation and covers the former method’s shortcomings. Unlike the BC transform, it involves the transformation of time series with positive and negative domains and has more potential for normalizing data.

\[
N_{\text{orm}}(\lambda) = \begin{cases} 
\frac{e^{\lambda DO(t)} - 1}{\lambda} & \lambda \neq 0 \\
DO(t) & \lambda = 0
\end{cases}
\]

\(N_{\text{orm}}(\lambda)\) normalized data. The values of the \(\lambda\), the transformation parameter, are \(-5 > \lambda > 5\).

Any predictable and repetitive pattern in time series results is evidence of nonstationarity. The stationary condition is the base assumption in stochastic modeling, which can be evaluated by various tests such as the augmented Dickey–Fuller test (ADF) [30] and autocorrelation function graphs. The stationarity evaluation in this test is done by fitting a regression line to the DO time series and determining the roots of the fitted equation. In this test, the stationary assumption declines when the corresponding probability to the test statistic \((P_{\text{ADF}})\) is higher than the significance level of 5%.

The evaluation of gradual oriented changes in DO time series can be assessed by the Mann–Kendall test [31]. This test can determine the gradual changes in time series that may occur alternately and seasonally, by considering both seasonal and nonseasonal parameters, leading to a seasonal or nonseasonal trend in the time series. The trendless DO time series assumption is rejected when the test’s statistic probability is higher than the significance level of 5%.

Steps in DO time series can be created by both human or natural factors and distinguished by sudden ascending or descending patterns in the time series. These sudden changes are easily observed from the raw values of the time series. A numerical method to identify these changes is the Mann–Whitney (MW) test [32]. This test evaluates the steps in the DO time series by ordering the data points and dividing them into subclasses and comparing the groups. A corresponding probability lower than the confidence level \(\alpha = 1\%\) rejects the step-less DO time series assumption.

One of the easiest and most used tools to check the periodicity in the DO time series is the correlogram (ACF). With the help of the correlogram, this property can visually be observed in the time series in the form of alternating variations in distinct and sinusoidal intervals. Alternatively, the numerical Fisher’s test [33] can be utilized to evaluate the periodicity in the DO time series. The test, similar to the spectral analysis, uses the Fourier coefficients to assess the significant periodic patterns. For the significance level of 5%, the degrees of freedom in the denominator of the test equal 3 and the test statistics higher than 3 depict periodicity in DO time series.

To study the techniques of water quality data preprocessing and their modeling methods, based on the flowchart in Figure 1, two scenarios are specified. In the first scenario, the data is only normalized, then, using different tests and differencing, the possibility of modeling with other stochastic models is considere. In the second scenario, the data are stationarized after normalization, and the feasibility of modeling with different linear methods is investigated. The modeling and preprocessing methods are compared based on various indices.

The indices used in evaluating DO time series include the conventional coefficient of determination \((R^2)\), variance accounted for (VAF), scatter index (SI), the root mean squared error (RMSE), and mean absolute error (MAE). A corrected Akaike’s information criterion (AICc) is used to choose a parsimonious ARIMA model. The Theil’s coefficients and Nash–Sutcliffe (EN-S) model efficiency coefficient [34–36] are used to assess the quality of modeling as well. The Theil test performance is based on the accuracy of forecasting (denoted as \(U_I\)) and forecasting quality (denoted as \(U_{II}\)). Using the AICc, \(U_I\) and \(U_{II}\) indices, superior models can be obtained in each time series modeling, and the less these indices are, the more parsimonious and more accurate the model is. Conversely, the higher the EN-S is, the better the model is.
Figure 1. Flowchart of preprocessing and modeling procedure.

In stochastic modeling, Ljung–Box (lbq) test [37] is one of the popular methods to investigate the model’s adequacy by testing models’ residuals. The corresponding calculated probability to the lbq test statistic in the $\chi^2$ distribution less than the $\alpha$ confidence level ($P_{lbq} > 5\%$), means that the model is not adequate and improvements in the preprocessing and/or modeling can take place, and better results can be obtained.

2.2. Study Site

The case study site to evaluate the ARIMA model is located in the Credit River Watershed, which is approximately 1000 km$^2$ in size and drains to Lake Ontario, Canada. The upper and middle areas of the watersheds are predominantly rural and include larger portions of the Town of Orangeville, Town of Erin, Town of Halton Hills, and Town of Caledon. The lower parts of the watershed, however, are mainly urbanized; and include the City of Brampton, Town of Oakville, and City of Mississauga. The spatial distribution of water quality trends in this basin is driven by land use, with the lower tributaries being impacted by urbanization and the resulting point and nonpoint pollution [38]. The dissolved oxygen (DO) data for this study was obtained from two water quality monitoring stations in the Credit River operated by the Credit Valley Conservation Authority in Mississauga, ON, Canada (Figure 2).

The first station (DO I) is situated at Old Derry Road at 43°37'19.1" N 79°44'00.1" W, and the second (DO II) is located 20.2 km downstream at the Mississauga Golf and Country Club (DO II) with the coordinates of 43°33'16.7" N 79°37'13.0" W. Water quality was monitored using a Hydrolab DSSX multiparameter sonde at each location. DO was measured using a Hatch luminescent dissolved oxygen sensor with an accuracy of ±0.1 mg/L at <8 mg/L and ±0.2 mg/L at >8 mg/L with a resolution of 0.01 mg/L. Sensor data were polled every 15 min and transferred via SODA telemetry to a central database. For DO I, the period of record is from 20 February 2010 to 11 December 2016 for 2337 days (221,819 data points, after removing errors and interpolation of missing data). For DO II, the period of record is from...
6 October 2011 to 11 December 2016 for 1865 days (177,819 data points). Figure 3 shows the seasonal fluctuations of the DO data during the 7 years of monitoring.

Figure 2. Credit River Watershed (reprinted from Credit Valley Conservation, [39]).

Figure 3. The plot of water quality dissolved oxygen (DO) index data for both stations.

3. Results and Discussions

3.1. Diurnal and Seasonal DO Fluctuations

In the preliminary investigation of the data, some missing data were observed. The missing data were reconstructed [40] by using the linear interpolation method. After data reconstruction, the data were averaged daily to obtain the daily average DO index. The data were divided into two
After normalization, data were analyzed by the JB test. It was found that the normalization transformation had little effect on the time series, and the value of the test statistics was not lower than the critical value, with the series close to the normal distribution. Therefore, the preprocessing steps of the time series are continued. In the next step, the time series were stationarized using the methods of standardization and spectral analysis. The results are shown in Table 2.

| Statistic | Nbr. | Min. | Max. | 1st Q. | Median | 3rd Q. | Mean | σ(n) | γ₁ | γ₂ |
|-----------|------|------|------|--------|--------|--------|------|------|----|----|
| DO I      | 2346 | 0.00 | 14.67| 8.80   | 10.53  | 12.68  | 10.68| 2.14 | −0.28| 0.39|
| Train     | 1642 | 6.47 | 14.67| 8.95   | 10.63  | 12.80  | 10.81| 2.06 | 0.07 | −1.32|
| Test      | 695  | 5.93 | 14.44| 8.55   | 10.35  | 12.29  | 10.48| 2.02 | 0.11 | −1.41|
| DO II     | 1865 | 5.63 | 14.42| 8.71   | 10.37  | 12.52  | 10.59| 2.01 | 0.11 | −1.36|
| Train     | 1306 | 6.88 | 14.42| 8.87   | 10.61  | 12.62  | 10.74| 1.96 | 0.07 | −1.40|
| Test      | 599  | 5.63 | 14.08| 8.37   | 9.75   | 12.20  | 10.26| 2.09 | 0.24 | −1.29|

Nbr., Number of data; Min. and Max., Minimum and Maximum of data; 1st Q. and 3rd Q., first and third Quarters; σ(n), Standard Deviation; γ₁, Skewness; γ₂, Kurtosis.

The time series was normalized using the Manly transformation to meet the modeling requirements. After normalization, data were analyzed by the JB test. It was found that the normalization transformation had little effect on the time series, and the value of the test statistics was not lower than the critical value, with the series close to the normal distribution. Therefore, the preprocessing steps of the time series are continued. In the next step, the time series were stationarized using the methods of standardization and spectral analysis. The results are shown in Table 2.

| Tests | Trend | Jump | Period | Stationary |
|-------|-------|------|--------|------------|
|       | MK (%)| SMK (%)| MW (%)| F * | ADF (%) |
| DO I  | 0.17  | 2.25  | 0.01  | 228 | 35.14 |
| Nnorm | 0.17  | 2.25  | 0.01  | 73  | 34.20 |
| S₁d   | 0.17  | 2.25  | 0.01  | −1 × 10⁹ | 34.20 |
| S₁    | 1.75  | 3.53  | 0.01  | 0   | 31.91 |
| DO II | 0.01  | 0.01  | 0.01  | 114 | 56.81 |
| Nnorm | 0.01  | 0.01  | 0.01  | 44  | 52.66 |
| S₁d   | 0.01  | 0.01  | 0.01  | −3 × 10⁹ | 52.66 |
| S₁    | 0.01  | 0.01  | 0.01  | 0   | 10.16 |

*. Fisher critical value: 3; augmented Dickey–Fuller (ADF) stationarity range: <5%.

The periodicity reduction in the ACF chart of time series can be observed in Figure 4. The spectral analysis method was able to significantly reduce the frequency of the DO II series, while the results of all these three preprocessing approaches are close for DO I. First-order differencing in the ARIMA model was also used to investigate the stationarizing impact of the model on the time series.

![Figure 4. Autocorrelation function plot of the preprocessed DO data.](image-url)
Table 3 demonstrates that all preprocessed time series are stationary and the probability of the ADF test statistic for all is below 5%. All deterministic terms in the series have been significantly reduced. The results of the tests applied to the normalized and standardized series are very similar for preprocessing methods at each station.

| Tests | Trend | Jump | Period | Stationary |
|-------|-------|------|--------|------------|
| Stats. | MK (%) | SMK (%) | MW (%) | (F *) * | ADF (%) |
| Norm | 70.15 | 46.68 | 55.20 | −0.24 | 0.01 |
| DO I | Std | 70.15 | 46.68 | 55.20 | −0.25 | 0.01 |
| | Sf | 71.87 | 35.53 | 53.10 | −0.19 | 0.01 |
| | Norm | 86.95 | 74.76 | 82.85 | −0.18 | 0.01 |
| DO II | Std | 86.95 | 54.76 | 82.85 | −0.18 | 0.01 |
| | Sf | 62.41 | 82.69 | 66.93 | −0.18 | 0.01 |

* Fisher critical value: 3; ADF stationarity range: <5%.

The spectral analysis method is slightly different from the other two methods. The complete elimination of the periodicity in the subtracted series is shown in the ACF graphs in Figure 5. In addition to eliminating seasonal fluctuations, nonseasonal correlations have also been eliminated, and preprocessed time series are completely stationary in the same first lag. The differencing resulted in improvement, from the values in Table 3, of MK (%) = 70.027, SMK (%) = 40.287, MW (%) = 54.490, and ADF (%) = 33.427 for the first station and MK (%) = 78.76, SMK (%) = 70.73, MW (%) = 77.53, and ADF (%) = 38.48 for the second station. The periodicity, though was increased in both stations from $-2.41 \times 10^9$ and $-7.11 \times 10^9$ to $-2.28 \times 10^{-1}$ and $-1.81 \times 10^{-1}$ for Stations I and II, respectively, the periodicity is insignificant in preprocessed time series. Due to the close results of statistical tests or correlograms, any specific preprocessing method cannot be introduced as a superior method in this stage.

![Figure 5. Autocorrelation function plot of DO data (ARIMA first-order differencing operator) at both DO stations I and II.](image)

3.2. Modeling Results

After the preprocesses using the correlograms method, the primary lag was completely damped. The maximum required modeling parameters for ARIMA modeling are of the first order.
However, methods to investigate the effect of increasing the order of parameters for all 10 parameters were considered. Thus, the parameters \( p = q = \{0, 1, 2, 3, ..., 8, 9, 10\} \) and the differencing order, \( d = 1 \) were considered. As a result, for each time series by varying the \( p \) and \( q \) parameters, 99 models for each case and in general for all six methods \( 6 \times 99 = 594 \) stochastic models were developed and reviewed. Series modeling results are presented in Table 4. As the results are very close to each other and the errors are too small, the values of Table 4 are multiplied by 100 for better presentation, except for the AICc index.

Table 4. Modeling evaluation indices of DO data.

| Superior Models | \( R^2 \) | VAF | RMSE | SI | MAE | \( E_{N-S} \) | AICc | \( U_I \) | \( U_H \) |
|-----------------|---------|-----|------|---|-----|----------|------|--------|--------|
| \( N_{orm} \)   | (4,1,4) | 0.943 | 0.943 | 0.482 | 4.601 | 0.353 | 0.943 | −997   | 0.0226 | 0.0452 |
| \( S_{ld} \)    | (2,1,4) | 0.943 | 0.943 | 0.483 | 4.607 | 0.354 | 0.943 | −999   | 0.0226 | 0.0452 |
| \( S_f \)       | (4,1,4) | 0.944 | 0.943 | 0.482 | 4.602 | 0.353 | 0.943 | −997   | 0.0226 | 0.0452 |
| \( N_{orm} \)   | (2,1,2) | 0.942 | 0.942 | 0.504 | 4.916 | 0.380 | 0.942 | −757   | 0.0241 | 0.0482 |
| \( S_{ld} \)    | (2,1,2) | 0.942 | 0.942 | 0.504 | 4.916 | 0.380 | 0.942 | −757   | 0.0241 | 0.0482 |
| \( S_f \)       | (6,1,0) | 0.941 | 0.941 | 0.506 | 4.954 | 0.362 | 0.941 | −748   | 0.0242 | 0.0484 |

* The RMSE and MAE values are in (mg/L).

It is observed that the coefficient of determination for all series is above 94% and very close to each other. Furthermore, error and modeling-power-survey indices are also very close. In the preprocessed series from the first station, the values of the model-surveying indices are almost identical for both \( N_{orm} \) and \( S_f \) methods, and with a slight difference, better than \( S_{ld} \). Therefore, considering fewer parameters (AICc = −999.57) of the \( S_{ld} \) method and consequently, simplicity of the model, \( S_{ld} \) is chosen as the superior preprocessing method for the DO I station.

In the time series modeling of the DO II station, the \( N_{orm} \) and \( S_{ld} \) methods that obtained close values are slightly better than \( S_f \) results. Considering the effect of the \( S_f \) method on reducing the periodicity in the DO II series in the ACF diagram of Figure 4, this method can be selected as a superior method with a slight difference from other methods in preprocessing, but with an insignificant difference, the \( S_{ld} \) model results were better than this method. The AICc index for \( S_{ld} \) equals −756.846, which is slightly better than the \( S_f \) method. Broadly, the \( S_{ld} \) can be chosen as the superior method of the second station as well.

Figures 6–8 examine the existence of a significant periodicity in model residuals and modeling accuracy [41,42]. Figure 6 shows the cumulative periodicity of the residuals of the ARIMA models. The cumulative residuals periodogram (CP) values of the modeled series are within 1% of the Kolmogorov–Smirnov confidence intervals. This means that the periodicity of the model residuals has been eliminated.

![Figure 6. Cumulative periodogram of residuals of ARIMA models for both DO stations I and II.](image_url)
The accuracy of the ARIMA stochastic models at the DO II station is well illustrated. These values are within the Chi-square 95% confidence ellipsis and close to the regression line.

This evaluation is performed to assess the adequacy of the stochastic models. In the case of having significant correlations in the residuals, the modeling process should be reviewed. In this figure, the results of the independence test for both stations’ data are presented. Using the daily data time series, the first 365 primary nonseasonal lags or the first primary seasonal lag of the residuals from the ARIMA models were evaluated by lbq test. Since the results of the three methods for each station are very close, the corresponding lbq test results are overlapped. For Station II (DO II) model residuals, total independence for the residuals is observed. In the second station, however, some negligible correlation is seen in the first lags for all three methods. This correlation might be due to periodicity in the DO I time series. Consequently, as the lbq values for the correlated lags are close to the significant 5% line, the correlations are removed immediately after few lags and the rest of the residual series are independent, it can be concluded that the models are adequate for DO I time series as well.

The decline in the independency is also seen in the initial lags of the model residuals for the DO II series, yet all the residuals of the models are independent. To study the modeling accuracy and correlation of the modeled values, scatter plots of the data are presented in Figure 8. The accuracy of the ARIMA stochastic models at the DO II station is well illustrated. These values are within the Chi-square 95% confidence ellipsis and close to the regression line.

However, at the DO I station, it is difficult to interpret due to outlier data. Therefore, another method is required, which is not affected by outliers. The outlier values can be caused by natural events or human
interactions in the watershed, which suddenly change the values of water quality indices. In addition, the existence of errors in the sensor measurements can also affect the creation of outlier data [43,44].

Figures 9 and 10 who the Taylor [45] diagram and forecast plots of the DO time series vs. linear stochastic models, respectively. The Taylor diagram compares the standard deviation centered root mean square difference (RMSD) and correlation coefficient (R²) simultaneously. Figure 9 shows that the model is accurately replicating the statistical characteristics of the original data [44–46].

![Figure 9. The Taylor diagram of the modeled DO time series at both stations I and II.](image1)

![Figure 10. The forecasts plot of the modeled DO time series at both DO Stations I and II for the superior method.](image2)

In the Taylor diagram of both stations, the proposed linear methodology gave very close results for both stations. Except that in the Station DO II, the $S_f$ method forecasted all DO data with lower standard deviation. Alongside other characteristics of the model, the stochastic model should preserve the statistical attributes of the original DO data [46]. Thus, Table 5 presents the descriptive statistics of the forecasted DO time series vs. the observed ones. The forecasted DO data closely match the observed ones in both stations. Figure 10 illustrates the superior model vs. the observed series. The capability of the one-step forward stochastic model in forecasting DO time series can be seen in the figure. The method correctly forecasted the fluctuations in the original series.
The DO time series follows a predictable diurnal cycle (Figure 11) which is driven by both water temperature and the mass balance between DO sources (re-aeration and production by photosynthesis) with DO consumption (by biochemical oxygen demand (BOD) and plant respiration). This cycle is closely tied to the solar cycle and can be modeled using a half-sinusoid function. [47,48]. The methods that we have applied to the daily average DO may also be used to predict minimum and maximum DO. The sub-hourly time series may then be reconstructed assuming the timing of the minimum and maximum DO values are fixed relative to sunset and sunrise.

![Figure 11. Relationship between diurnal DO cycle, DO daily average, and estimated DO daily minimum based on linear correlation.](image)

For the protection of aquatic life, the minimal DO values are the most important for both acute and chronic impacts. The daily minimum DO values are linearly correlated with the daily average DO, having a Pearson Correlation Coefficient of 0.98 for both DO I and DO II. This correlation is shown in Figure 11, where the estimated daily minimum based on linear regression is shown compared to the original time series and daily average. Comprehensive DO water quality metrics focus on preventing both the acute and chronic effects by considering a lower limit on instantaneous values, as well as limits based on a long-term average [49–53]. The average seven-day minimum is a commonly used metric for chronic exposure of low DO [50,52,53] and can be simply calculated based on the above-mentioned linear relationship between the daily average and daily minimum. Using this method produces results with acceptable error during the summer period between June and September when the lowest DO values occur seasonally (RMSE = 0.41 mg/L, MAE = 0.31 mg/L for DO I, RMSE = 0.37 mg/L, MAE = 0.29 mg/L for DO II).

### 4. Conclusions

In this study, the dissolved oxygen (DO) time series of two stations at different geographical points along the Credit River was used to examine the preprocessing and one-step-ahead stochastic
modeling techniques. The data were measured in 15 min intervals at two different time periods. Initially, using the linear interpolation method, missing data were reconstructed and then averaged daily to obtain the daily DO values.

The initial premises of stochastic methods are the normal distribution of the time series and their stationarity. Jarque–Bera (JB) test was used to determine the distribution of the series. The stationarity of both series was studied in two steps. In the first step, the general stationarity was studied using the augmented Dickey–Fuller (ADF) test. In the second step, the Mann–Whitney, Mann–Kendall, and Fisher tests were used to examine the factors causing the nonstationarity of the series. It was observed that both time series are nonstationary and non-normally distributed. The nonstationarity of the time series is due to the existence of a trend, jump, and periodicity. After scrutinizing the test results, the ACF and PACF graphs were considered for modeling. In the first scenario, the data were normalized and then, by using the first-order differencing operator in ARIMA, they were stationarized.

In the second scenario, the series was stationarized after normalization and then modeled with the ARIMA. Initially, daily data were normalized by Manly transforms (Norm). This transformation was unable to normalize the distribution of time series but brought the distribution closer to the normal distribution in both time series. Then, the values were stationarized by the methods of differencing, standardization (Std), and spectral analysis (Sf). It was found that the two methods, Sd and Sf, which were used to stationarize the series after normalization, are only able to reduce the periodicity in the series and practically cannot stationarize them alone. However, using the differencing operator in the stochastic ARIMA model, the preprocessed series were stationarized. According to the results of numerous tests, values of the stationarity and the elimination of deterministic terms are very close to each other. After modeling the preprocessed time series, using the stochastic ARIMA model, the results showed that the methods performed similarly to each other and were very accurate. The Sd method was the superior method of the time series preprocessor for the first and second stations outperformed other methods.

Dissolved oxygen is one of many important water quality parameters that are measured at real-time water quality stations and is a crucial parameter for assessing stream habitat suitability. This preprocessing approach can be used for other water quality parameters such as temperature, which exhibit similar periodicity and trends and applied to other models. For example, a novel genetic algorithm (GA)-optimized long short-term memory (LSTM) water temperature model developed by Stajkowski et al. [4] can be combined with this technique to increase accuracy. The main goal of developing accurate models of key water quality parameters is the forecasting and assessment of the impact of disturbances (anthropogenic, such as pollution, or climatic such as climate change) on aquatic habitat suitability and, therefore, the health of aquatic species. There is a strong linear relationship between daily average DO and daily minimum DO, which allows for water quality metrics such as the average seven-day minimum DO to be predicted from this model.

It can be finally concluded that with the correct selection of preprocessing and stochastic methods, which are reliable, simple, and well understood, it is possible to accurately model water quality indices. This is in contrast to artificial intelligence methods, which are a black box and where the connection between model structure and the time series components (both deterministic and stochastic) are not explicit. Additionally, the preprocessing methods used in this study can be applied in combination with artificial intelligence models, including hybrid models such as ARIMA–ANN, ARIMA–long short-term memory, and ARIMA–ELM to improve forecasting accuracy. Moreover, by utilizing other stationarity methods like the advanced exponential smoothing method, the impacts of these methods on DO time series preprocessing can be investigated.

**Author Contributions:** Conceptualization, S.S. and B.G.; methodology, S.S. and M.Z.; software, M.Z.; validation, S.S., H.B.; formal analysis, S.S. and H.B.; resources, S.S., B.G.; data curation, S.S.; writing—original draft preparation, S.S. and M.Z.; writing—review and editing, S.S., H.F., B.G. and H.B.; visualization, S.S. and M.Z.; supervision, H.F., B.G. and H.B.; funding acquisition, H.F. and B.G. All authors have read and agreed to the published version of the manuscript.
**Funding:** This research was funded by an NSERC Discovery Grant (#400675), and in partnership with the Ontario Ministry of Transportation Grant (#050235).

**Acknowledgments:** River temperature data were generously provided by the Credit Valley Conservation Authority.

**Conflicts of Interest:** The authors declare no conflict of interest for the current work.

**References**

1. Antanasijević, D.; Pocajt, V.; Perić-Grujić, A.; Ristić, M. Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis. *J. Hydrol.* **2014**, *519*, 1895–1907. [CrossRef]
2. Bayram, A.; Uzlu, E.; Kankal, M.; Dede, T. Modeling stream dissolved oxygen concentration using teaching–learning based optimization algorithm. *Environ. Earth Sci.* **2015**, *73*, 6565–6576. [CrossRef]
3. Rajaee, T.; Khani, S.; Ravansalar, M. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103978. [CrossRef]
4. Staikowski, S.; Kumar, D.; Samui, P.; Bonakdari, H.; Gharabaghi, B. Genetic-Algorithm-Optimized Sequential Model for Water Temperature Prediction. *Sustainability* **2020**, *12*, 5374. [CrossRef]
5. Bonakdari, H.; Binns, A.D.; Gharabaghi, B. A Comparative Study of Linear Stochastic with Nonlinear Daily River Discharge Forecast Models. *Water Resour. Manag.* **2020**, *34*, 3689–3708. [CrossRef]
6. Cox, B.A. A review of dissolved oxygen modelling techniques for lowland rivers. *Sci. Total Environ.* **2003**, *314*, 303–334. [CrossRef]
7. Feaster, T.D.; Conrads, P.A. Characterization of Water Quality and Simulation of Temperature, Nutrients, Biochemical Oxygen Demand, and Dissolved Oxygen in the Wateree River, South. Carolina, 1996–1998; U.S. Department of the Interior, US Geological Survey: Columbia, SC, USA, 2000.
8. Huang, J.; Yin, H.; Chapra, S.C.; Zhou, Q. Modelling Dissolved Oxygen Depression in an Urban River in China. *Water* **2017**, *9*, 520. [CrossRef]
9. Schmidt, A.R.; Stamer, J.K. Assessment of Water Quality and Factors Affecting Dissolved Oxygen in the Sangamon River, Decatur to Riverton, Illinois, Summer 1982; U.S. Department of the Interior, US Geological Survey: Urbana, IL, USA, 1987.
10. Waldron, M.C.; Wiley, J.B. *Water Quality and Processes Affecting Dissolved Oxygen Concentrations in the Blackwater River, Canaan Valley, West. Virginia*; U.S. Department of the Interior, US Geological Survey: Charleston, WV, USA, 1996.
11. Langridge, M.; Gharabaghi, B.; McBean, E.; Bonakdari, H.; Walton, R. Understanding the dynamic nature of Time-to-Peak in UK streams. *J. Hydrol.* **2020**, *583*, 124630. [CrossRef]
12. Zaji, A.H.; Bonakdari, H.; Gharabaghi, B. Developing an AI-based method for river discharge forecasting using satellite signals. *Theor. Appl. Climatol.* **2019**, *138*, 347–362. [CrossRef]
13. Zounemat-Kermani, M.; Seo, Y.; Kim, S.; Ghorbani, M.A.; Samadianfard, S.; Naghshara, S.; Kim, N.W.; Singh, V.P. Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida. *Appl. Sci.* **2019**, *9*, 2534. [CrossRef]
14. Sentas, A.; Psilovikos, A.; Psilovikos, T.; Matzaferti, N. Comparison of the performance of stochastic models in forecasting daily dissolved oxygen data in dam-Lake Thesaurus. *Desalin. Water Treat.* **2016**, *57*, 11660–11674. [CrossRef]
15. Heddam, S.; Kisi, O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol.* **2018**, *559*, 499–509. [CrossRef]
16. Harvey, R.; Lye, L.; Khan, A.; Paterson, R. The influence of air temperature on water temperature and the concentration of dissolved oxygen in Newfoundland Rivers. *Can. Water Resour. J.* **2011**, *36*, 171–192. [CrossRef]
17. Bertone, E.; Stewart, R.A.; Zhang, H.; Veal, C. Data-driven recursive input–output multivariate statistical forecasting model: Case of DO concentration prediction in Advancetown Lake, Australia. *J. Hydroinformatics* **2015**, *17*, 817–833. [CrossRef]
18. Parmar, K.S.; Bhardwaj, R. Water quality management using statistical analysis and time-series prediction model. *Appl. Water Sci.* **2014**, *4*, 425–434. [CrossRef]
19. Heddam, S. Use of optimally pruned extreme learning machine (OP-ELM) in forecasting dissolved oxygen concentration (DO) several hours in advance: A case study from the Klamath River, Oregon, USA. *Environ. Process.* **2016**, *3*, 909–937. [CrossRef]

20. Chen, Y.; Xu, J.; Yu, H.; Zhen, Z.; Li, D. Three-dimensional short-term prediction model of dissolved oxygen content based on pso-bpnn algorithm coupled with kriging interpolation. *Math. Probl. Eng.* **2016**, *2016*, [CrossRef]

21. Chang, F.-J.; Chen, P.-A.; Liu, C.-W.; Liao, V.H.-C.; Liao, C.-M. Regional estimation of groundwater arsenic concentrations through systematical dynamical-neural modeling. *J. Hydrol.* **2013**, *499*, 265–274. [CrossRef]

22. Kisi, O.; Akbari, N.; Sanatipour, M.; Hashemi, A.; Teimourzadeh, K.; Shiri, J. Modeling of Dissolved Oxygen in River Water Using Artificial Intelligence Techniques. *J. Environ. Inform.* **2013**, *22*. [CrossRef]

23. Salas, J.D.; Yevjevich, V.; Lane, W.L.; Delleur, J.W. Applied Modeling of Hydrologic Time Series; Water Resources Publications: Littleton, CO, USA, 1980.

24. Ebtehaj, I.; Bonakdari, H.; Zeynoddin, M.; Gharabaghi, B.; Azari, A. Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models. *Int. J. Environ. Sci. Technol.* **2020**, *17*, 505–524. [CrossRef]

25. Zeynoddin, M.; Bonakdari, H.; Ebtehaj, I.; Esmaeilbeiki, F.; Gharabaghi, B.; Haghi, D.Z. A reliable linear stochastic daily soil temperature forecast model. *Soil Tillage Res.* **2019**, *189*, 73–87. [CrossRef]

26. Moeeni, H.; Bonakdari, H.; Fatemi, S.E. Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. *J. Hydrol.* **2017**, *547*, 348–364. [CrossRef]

27. Lotfi, K.; Bonakdari, H.; Ebtehaj, I.; Mjalli, F.S.; Zeynoddin, M.; Delatolla, R.; Gharabaghi, B. Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology. *J. Environ. Manag.* **2019**, *240*, 463–474. [CrossRef]

28. Manly, B.F.J. Exponential Data Transformations. *J. R. Stat. Soc. Ser. D* **1984**, *33*, 100–105. [CrossRef]

29. Said, S.E.; Dickey, D.A. Testing for unit roots in autoregressive-moving average models of unknown order. *J. Forecasting* **1980**, *9*, 377–390. [CrossRef]

30. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B* **1964**, *26*, 211–243. [CrossRef]

31. Jain, A.; Kumar, A.M. Hybrid neural network models for hydrologic time series forecasting. *Appl. Soft Comput.* **2007**, *7*, 585–592. [CrossRef]

32. Zhang, S.; Zhou, Q.; Xu, D.; Lin, J.; Cheng, S.; Wu, Z. Effects of sediment dredging on water quality and zooplankton community structure in a shallow of eutrophic lake. *J. Environ. Sci.* **2010**, *22*, 218–224. [CrossRef]

33. Wichert, S.; Fokianos, K.; Strimmer, K. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* **2004**, *20*, 5–20. [CrossRef]

34. Burnham, K.P.; Anderson, D.R. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA; London, UK, 2002.

35. Theil, H. *Applied Economic Forecasting*; North-Holland Publishing Company: Amsterdam, The Netherlands, 1966.

36. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]

37. Kandel, S.; Parikh, R.; Paepecke, A.; Hellerstein, J.M.; Heer, J. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 547–554. [CrossRef]

38. Chang, H. Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Res.* **2008**, *42*, 3285–3304. [CrossRef] [PubMed]

39. Credit Valley Conservation (CVC). Watershed Monitoring: Real-Time Water Quality. Available online: [http://www.creditvalleyca.ca/watershed-science/watershed-monitoring/real-time-water-quality/](http://www.creditvalleyca.ca/watershed-science/watershed-monitoring/real-time-water-quality/) (accessed on 6 April 2017).

40. Gnauck, A.; Luther, B. Missing Data in Environmental Time Series--a Problem Analysis. In *EnviroInfo*; Masaryk University: Brno, Czech Republic, 2005; pp. 848–852. ISBN 80-210-3780-6.

41. Jayawardena, A.W.; Lai, F. Time series analysis of water quality data in Pearl River, China. *J. Environ. Eng.* **1999**, *115*, 590–607. [CrossRef]

42. El-Din, A.G.; Smith, D.W. A combined transfer-function noise model to predict the dynamic behavior of a full-scale primary sedimentation tank. *Water Res.* **2002**, *36*, 3747–3764. [CrossRef]

43. Faruk, D.O. A hybrid neural network and ARIMA model for water quality time series prediction. *Eng. Appl. Artif. Intell.* **2010**, *23*, 586–594. [CrossRef]
44. Antonopoulos, V.Z.; Papamichail, D.M.; Mitsiou, K.A. Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. *Hydrol. Earth Syst. Sci.* **2001**. [CrossRef]

45. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **2001**, *106*, 7183–7192. [CrossRef]

46. Singh, V.P.; Frevert, D.K. *Mathematical Models of Small Watershed Hydrology and Applications*; Water Resources Publication: Highlands Ranch, CO, USA, 2002.

47. Dobbins, W.E. BOD and oxygen relationship in streams. *J. Sanit. Eng. Div.* **1964**, *90*, 53–78.

48. Chapra, S.C.; Di Toro, D.M. Delta method for estimating primary production, respiration, and reaeration in streams. *J. Environ. Eng.* **1991**, *117*, 640–655. [CrossRef]

49. Davis, J.C. Minimal dissolved oxygen requirements of aquatic life with emphasis on Canadian species: A review. *J. Fish. Board Can.* **1975**, *32*, 2295–2332. [CrossRef]

50. Chapman, G. *Ambient Water Quality Criteria for Dissolved Oxygen*; US EPA: Washington, DC, USA, 1986; EPA 440/5-864/03.

51. Canadian Council of the Ministers of Environment (CCME). *Canadian Water Quality Guidelines for the Protection of Aquatic Life: Dissolved Oxygen*; Canadian Council of Ministers of the Environment: Winnipeg, MB, Canada, 1999.

52. Grand River Water Management Plan. *Water Quality Targets to Support. Healthy and Resilient Aquatic Ecosystems in the Grand River Watershed*; Prepared by the Water Quality Working Group; Grand River Conservation Authority: Cambridge, ON, Canada, 2013.

53. Franklin, P.A. Dissolved oxygen criteria for freshwater fish in New Zealand: A revised approach. *N. Z. J. Mar. Freshw. Res.* **2014**, *48*, 112–126. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).